UNIVERSITY OF VERONA

DEPARTMENT OF

Biotechnology

PHD SCHOOL

Natural Sciences and Engineering

PHD IN

Biotechnology

With funding by

University of Verona

CYCLE / YEAR of initial enrolment **29/2014**

PHD THESIS TITLE

Development and Application of Computational Biology tools for Biomedicine

S.S.D. **BIO10** (Biochemistry)

Coordinator: **Prof. Massimo Delledonne**

Tutor: **Prof. Alejandro Giorgetti**

PhD candidate: **Dr. Mirko Busato**

January 2017

Sommario

La simulazione biomolecolare può essere considerata come un microscopio virtuale per la biologia molecolare, permettendo di ottenere informazioni sui meccanismi subcellulari di rilevanza biologica a scale spaziali e temporali che sono difficili da osservare sperimentalmente. Essa fornisce un potente strumento per collegare le leggi della fisica con il complesso comportamento dei sistemi biologici.
Notevoli e recenti miglioramenti in termini di velocità di simulazione raggiungibile e dei modelli fisici sottostanti porteranno sempre più ad investigazioni molecolari di grandi sistemi. Questi miglioramenti possono per gran parte influenzare le scienze biologiche.
In questa tesi, ho applicato approcci di biologia molecolare computazionale per diversi sistemi biologici utilizzando strumenti di bioinformatica strutturale e biofisica computazionale allo stato dell'arte. Il mio obiettivo principale era il design computazionale dei polimeri a stampo molecolare (MIPs) che hanno recentemente attirato grande attenzione come convenienti sostituti per gli anticorpi naturali e recettori in cromatografia, sensori e saggi. Ho usato la modellizzazione molecolare per ottimizzare le composizioni polimeriche per produrre recettori sintetici ad alta affinità basati sull' imprinting molecolare.
In particolare ho sviluppato un protocollo nuovo e gratis che può essere eseguito nel giro di poche ore e che genera un elenco di candidati monomeri che sono in grado di formare forti interazioni di legame con il modello. Inoltre, ho prodotto un nuovo metodo computazionale per il calcolo dei rapporti stechiometrici tra monomero e molecola templante, da utilizzare come punto di partenza in laboratorio per la sintesi dei MIPs. Questi protocolli sono stati implementati in un web server che è disponibile all'indirizzo: http://mirate.di.univr.it/.
In parallelo, ho studiato anche la modellazione di sistemi di MIPs molto più complessi, introducendo alcuni fattori come le molecole di solvente e cross-linker, le quali sono essenziali nel processo di polimerizzazione. Un nuovo algoritmo, che imita il meccanismo di polimerizzazione radicalica, è stato scritto per essere applicato nella progettazione razionale dei MIPs.
Sono stato inoltre coinvolto nel campo della medicina biomolecolare computazionale. Infatti, nei capitoli 5 e 6 descrivo il lavoro svolto in collaborazione con due laboratori del dipartimento di Neuroscience, Biomedicina e Movimento, all'Università di Verona. Nel capitolo 5, partendo da dati sperimentali non pubblicati, ho caratterizzato computazionalmente l'interazione dell'ACOT8 con Nef di HIV-1. Ho effettuato una caratterizzazione strutturale e funzionale dettagliata di queste due proteine al fine di dedurre eventuali dettagli funzionali sulle loro interazioni. Le predizioni bioinformatiche sono stati poi confermate da esperimenti di wet-lab.
Ho effettuato anche una caratterizzazione strutturale e funzionale dettagliata di due mutazioni patogene di AGT-Mi (capitolo 6). In particolare, ho usato simulazioni di dinamica molecolare classica per studiare le possibili interferenze con il processo di dimerizzazione di AGT-Mi esercitate dai mutanti I244T-Mi e F152I-Mi. Queste varianti sono associate con la malattia Primary Hyperoxaluria type 1. Nel capitolo 7, presento simulazioni di dinamica molecolare coarse-grain delle interazioni tra la membrana e la proteina Human ileal bile-acid-binding. Questo studio è stato condotto in collaborazione con il gruppo di NMR dell'Università di Verona e fa parte di una vasta ricerca finalizzata alla migliore comprensione delle principali interazioni biomolecolari in ambienti cellulari

crowded. I risultati delle simulazioni di dinamica molecolari erano in accordo con i risultati sperimentali.

Abstract

Biomolecular simulation can be considered as a virtual microscope for molecular biology, allowing to gain insights into the sub-cellular mechanisms of biological relevance at spatial and temporal scales that are difficult to observe experimentally. It provides a powerful tool to link the laws of physics with the complex behavior of biological systems. Dramatic recent advancements in achievable simulation speed and the underlying physical models will increasingly lead to molecular views of large systems. These improvements may largely affect biological sciences.
In this thesis, I have applied computational molecular biology approaches to different biological systems using state of the art structural bioinformatics and computational biophysics tools (Chapter 3).
My principal focus was on the computational design of molecular imprinted polymers (MIPs), which have recently attracted significant attention as cost effective substitutes for natural antibodies and receptors in chromatography, sensors and assays. I have used molecular modelling in the optimization of polymer compositions to make high affinity synthetic receptors based on Molecular Imprinting. In particular, I developed a new free of charge protocol that can be performed within just few hours that outputs a list of candidate monomers which are capable of strong binding interactions with the template. Furthermore, I have produced a new computational method for the calculation of the ideal monomer: template stoichiometric ratio to be used in the lab for the MIPs synthesis. These protocols have been implemented as a webserver that is available at http://mirate.di.univr.it/.
In parallel, I have also investigated the modelling of much more complex MIPs systems by the introduction of some factors e.g. solvent and cross-linker molecules that are also essential in the polymerisation process. A novel algorithm, which mimics a radical polymerization mechanism, has been written for application in the rational design of MIPs (Chapter 4).
Moreover, I have been involved in the field of computational molecular biomedicine. Indeed, in Chapters 5 and 6 I describe the work done in collaboration with two labs at the Department of Neuroscience, Biomedicine and Movement Sciences, University of Verona. In Chapter 5, starting from unpublished experimental data I have computationally characterized the interaction of ACOT8 with HIV-1 Nef accessory protein. I have performed a detailed structural and functional characterization of these two proteins in order to infer any possible functional details about their interactions. The bioinformatics predictions were then confirmed by wet-lab experiments.
I have also carried out a detailed structural and functional characterization of two pathogenic mutations of AGT-Mi (Chapter 6). In particular, I have used classical molecular dynamics (MD) simulations to study the possible interference with the dimerization process of AGT-Mi exerted by I244T-Mi and F152I-Mi mutants. Those variants are associated with Primary Hyperoxaluria type 1 disease.

In Chapter 7, I present the coarse-grained MD simulations of Membrane/Human ileal bile-acid-binding protein Interactions. This study was carried out in collaboration with the NMR group at the University of Verona and it is a part of an extensive research aimed at better understanding of the main biomolecular interactions in crowded cellular environments. MD simulations results were in agreement with experimental findings.

TABLE OF CONTENTS

Sommario **3**
Abstract **4**

Chapter 1. Introduction **9**
1.1. Introduction to Computational Molecular Biology **9**
1.2 Computational Molecular Biology: Achievements and Challenges **9**

Chapter 2. Aims of this thesis **11**

Chapter 3. Computational Molecular Biology Methods **12**
3.1 Homology modeling **12**
3.1.1 Principle of homology modeling: advancements and limitations **12**
3.1.2 Steps in model generation **13**
3.1.2.1 Template search and target-template sequence alignment **14**
3.1.2.2 Template Selection **15**
3.1.2.3 Model construction and refinement **16**
3.1.2.4 Model quality assessment **17**
3.2 Molecular Docking **18**
3.2.1 Basic concepts of molecular docking **18**
3.2.2 Information-driven docking with HADDOCK **19**
3.2.3 Autodock suite **22**
3.2.4 LeapFrog (Sybyl) **24**
3.3 Molecular Dynamics **25**
3.3.1 Choice of simulation approach: length and time scales **25**
3.3.2 Molecular dynamics: introduction **26**
3.3.3 Periodic boundary conditions **27**
3.3.4 Force fields **28**
3.3.4.1 Coarse-grain methods **29**
3.3.4.1.1 MARTINI **29**
3.3.5 Solvation **30**
3.3.6 Cut-off **30**
3.3.7 Integration of the equations of motion **31**
3.3.8 Thermostats **31**

Chapter 4. Computational Design and Optimization of Molecularly Imprinted Polymers **33**
4.1 Introduction on Molecularly imprinted polymers (MIPs) **33**
4.1.2 General applications of MIPs **34**
4.1.3 Current methods of MIPs production **35**
4.1.3.1 Bulk Polymerization **35**
4.1.3.2 Precipitation Polymerization **36**
4.1.3.3 Grafting polymerization on monodisperse polymer beads **36**
4.1.3.4 Emulsion synthesis **36**
4.1.4 Rational Design of MIPs **37**
4.1.4.1 Combinatorial approach **37**
4.1.4.2 Thermodynamic approach **37**
4.1.4.3 Computational Design using Sybyl protocol **38**
4.2 Aims of the research described in this Chapter **39**
4.3 Development of an innovative web server aimed at the rational design of

MIPs **39**
4.3.1 Workflow **39**
4.3.1.1 Input preparation **41**
4.3.1.2 Docking step **42**
4.3.1.3 Refining step **43**
4.3.2 Case of study and Applications **43**
4.3.2.1 HADDOCK based docking step **43**
4.3.2.1.1 Troponin system **43**
4.3.2.1.1.1 Biological Introduction **43**
4.3.2.1.1.2 Aim of this section **44**
4.3.2.1.1.3 Materials and Methods **44**
4.3.2.1.1.4 Results **45**
4.3.2.1.1.4.1 Experiments on NK11 **46**
4.3.2.1.1.4.2 Experiments on NR11 **48**
4.3.2.1.1.4.3 Experiments on AK9 **50**
4.3.2.1.1.5 Discussion and Conclusions **52**
4.3.2.1.2 Hepcidin system **53**
4.3.2.1.2.1 Biological Introduction **53**
4.3.2.1.2.2 Aim of this section **54**
4.3.2.1.2.3 Materials and Methods **54**
4.3.2.1.2.4 Results **55**
4.3.2.1.2.4.1 Hepcidin: first test **56**
4.3.2.1.2.4.2 Hepcidin: second test **58**
4.3.2.1.2.5 Discussion and Conclusions **60**
4.3.2.2 Autodock/Vina based docking step **60**
4.3.2.3 GROMACS based refinement step **61**
4.3.2.3.1 Melamine **61**
4.3.2.3.1.1 Biological Introduction **61**
4.3.2.3.1.2 Aim of this section **62**
4.3.2.3.1.3 Materials and Methods **62**
4.3.2.3.1.4 Results and Discussion **63**
4.4 Implementation of a new algorithm to *in silico* synthesis of MIPs **65**
4.4.1 Introduction **65**
4.4.2 Materials and Methods **66**
4.4.3 Results and Discussion **69**
4.4.4 Conclusions **72**

Chapter 5. Molecular characterization of HIV-1 Nef and ACOT8 interaction: insights from *in silico* structural predictions **73**
5.1 Introduction **73**
5.2 Aims of the research described in this Chapter **74**
5.3 Materials and Methods **75**
5.4 Results **76**
5.4.1 ACOT8 homology modeling **76**
5.4.2 Initial model of Nef/ACOT8 interaction **77**
5.4.3 Knowledge-based Nef/ACOT8 complex model **80**
5.4.4 Comparison between Nef-LAI and Nef-SF2 **81**
5.5 Discussion and Conclusions **82**

Chapter 6. Molecular dynamics simulations of F152I-Mi and I244T-Mi variants associated with PH1 and implications in their pathogenicity **84**
6.1 Introduction **84**
6.2 Aims of the research described in this Chapter **85**

6.3 Material and Methods **85**
6.4 Results **87**
6.4.1 I244T-Mi mutant **87**
6.4.2 F152I-Mi mutant **88**
6.5 Discussion and Conclusions **89**

Chapter 7. Modeling human ileal bile-acid-binding protein/membrane interactions within a crowded cellular environment **91**
7.1 Introduction **91**
7.2 Aims of the research described in this Chapter **94**
7.3 Material and Methods **94**
7.4 Results **95**
7.5 Discussion and Conclusions **98**

Chapter 8. Conclusions **100**

Papers published during my PhD period **103**
References **104**
Appendix **126**

Chapter 1

Introduction

1.1 Introduction to Computational Molecular Biology

Structural bioinformatics, originally called "structural computational biology", precedes other forms of bioinformatics. Indeed, the James Watson and Francis Crick's classic article (Watson and Crick, 1953) can be actually considered a modeling paper and probably the first structural bioinformatics paper (Samish et. al, 2015). The 2013 Nobel Prize won by the Martin Karplus, Michael Levitt and Arieh Warshel for the development of multiscale models for complex chemical systems defines a significant hallmark, "certifying" the importance and the influence of structural bioinformatics on science. In "The birth of computational structural biology" article (Levitt, 2001), Prof. Levitt highlighted the importance and the need of computation approaches to correctly refine the tRNA structural model proposed by Crick. Hence, since the very beginning, computation has been an indispensable part of structural biology and over the time, it has spread to biochemistry and molecular biology area.
From the first short simulations of few picoseconds of few atoms, now millisecond simulations (Beauchamp et al., 2011) or extensive searches of sequence and structure space have become feasible tasks.
As stated in the Samish et al., 2014 paper, the expression 'structural bioinformatics' describes data-driven statistical, knowledge-based study of significant non-redundant cluster of structures in order to comprehend the system of interest.
On the other hand, 'computational biophysics' delineates a hypothesis-driven physics-based treatment of biomolecular systems. The ergodic hypothesis ensures that outcomes from the two methodologies converge over sizable non-redundant samples or extensive simulations (Samish, 2009).
"Computational Molecular Biology" (CMB), is a new discipline that brings together these two main approaches i.e. "structural bioinformatics and "computational biophysic". CMB is used to understand the function of biological macromolecules, the factors of specificity and selectivity in molecular interactions and the dynamic aspects of macromolecular structures and their influence on function and stability.

1.2 Computational Molecular Biology: Achievements and Challenges

In the last 10 years, the CMB field has seen several achievements and a dramatic growth in terms of data coverage and computational resources. World-wide genome-sequencing projects have made available a large amounts of sequence data and the number of three-dimensional (3D) structures deposited in PDB database has grown 4-fold to over 100,000 structures (Sillitoe et al., 2012). The increased availability of computer power has made possible the design of new dedicated supercomputer e.g. ANTON (Shaw et al., 2009) for molecular dynamics (MD) simulations. In the other hand, improvements in the field of structural bioinformatics were assessed using an objective method. Indeed, the Critical

assessment of protein structure prediction (CASP) editions have enabled important developments in the methods for the generation of 3D structural models, and recently, the CASP blind-prediction spirit has been extended to different levels e.g. protein interactions (Janin et al., 2003), function prediction (Radivojac et al., 2013), membrane protein docking (Kufareva et al., 2014). Moreover, new approaches, involving the use of correlated mutation data, have allowed the generation of pairwise amino acid contact maps from sequence data (Marks et al., 2011; Nugent and Jones, 2012). These contact maps have been exploited as spatial restraints to generate structural predictions. Furthermore, novel techniques that combine the 'omics' data with structural approaches e.g. the detection of binding region similarities have been adopted for the drug discovery and in the early phases of drug design.

Despite noteworthy advancements have been made; improvements are still needed in several areas. Modeling large or multi-domain proteins and assemblies such as ribosome (Yonath, 2010) and the proteasome can be sometimes a difficult task, as well as the modeling 3D RNA structures is still in its infancy. Indeed, although various computational techniques e.g. MD simulations and elastic network models can be useful to understand large conformational changes due to ligand/protein binding, allosteric effects or post-translational modifications, sometimes they present some challenges for complex systems (Najmanovich et al., 2008).

Another important point is the integration of the structural bioinformatics with the system biology. The latter is important for two main reasons: i) to have a full computationally comprehension of a living cell at all scales and ii) to recognize the functional factors of a biomolecule in order to predict if small differences (mutations) may have a drastic effect on the function at the molecular level. Some important questions about the origins and evolution of protein structure and the protein fold should be discussed; among which: "Is it possible to predict the structure of a protein based solely on its protein primary sequence?"; "why are some folds so much more common than others?". Although these questions are partially addresses they don't have a clear answer.

Finally, the importance of the accessibility and integration of data and methods in this field must not be overlooked. Indeed, the main success in this field, comes from the high number of efficient open source programs for different applications such as GROMACS (Pronk et al.,2013) for molecular dynamics simulations; HADDOCK (Dominguez et. al., 2003 and van Zundert et. al., 2016) and Autodock (Morris et. al., 2009) for molecular docking and MODELLER (Eswar et al., 2008) for homology modelling.

One challenge for the future is to make the existing methods accessible to newcomers in the field and to the scientific community in general. With this in mind, new dedicated databases and web-servers have been developed (Brazas et al., 2012 and Busato and Giorgetti, 2016).

Chapter 2

Aims of this thesis

In this thesis, I have developed and applied state of the art computational molecular biology approaches (Chapter 3) to gain insights into different biological systems of outstanding interest in three different fields of biotechnology: i) molecular imprinted polymer technology, ii) computational molecular biomedicine and iii) macromolecular crowding in living cell.

In the main part of my work (Chapter 4) the aim is to design protocols for the predictive rational design of Molecularly Imprinted Polymers (MIPs) for practical applications. The first part of Chapter 4 includes the development of a publicly accessible web server (MIRATE web server), which relies on novel free of charge protocols/methodologies implemetented by me and my collaborators, aimed at the intelligent design of MIPs. In the second part of Chapter 4, I present a new algorithm, written in Sybyl programming language, which mimics a radical polymerization mechanism.

My work in the field of computational molecular biomedicine is described in Chapters 5 and 6. In Chapter 5 I describe the work done in collaboration with the lab of Prof. D. Zipeto, in which, starting from HIV-1 Nef experimental data I have computationally characterized the human peroxisomal thioesterase 8 (ACOT8) and its interaction with with HIV-1 Nef. I have performed a detailed structural and functional characterization of these two proteins in order to infer any possible details about their interactions. The predictions were then validated by wet-lab experiments.

In Chapter 6, I present the research carried out in association with the lab of Prof. C. Voltattorni. In particular, I investigated the effects of interface mutations on the dimerization of alanine glyoxylate aminotransferase (AGT) and implications in the mistargeting of the pathogenic variants F152I and I244T by performing extensive classical, explicit-solvent MD simulations.

In Chapter 7, I used state of the art coarse-grained computational tools to gain a deeper understanding of the IBAP-anionic lipid membranes interactions, mimicking the macromolecular and membrane cellular components in a crowded cellular environment. This study was realized in collaboration with the NMR group led by Prof M. Assfalg.

Chapter 3

Computational Molecular Biology Methods

3.1 Homology Modeling

This chapter offers an overview on the computational methods used in the thesis. However, all computational details on the particular methods and the simulation setup for each system can be found in their corresponding chapters.

3.1.1 Principle of homology modeling: advancements and limitations

Homology or comparative modeling methods are based on the idea that evolutionary related proteins share a similar structure (Chothia and Lesk, 1986).
The quality of the predicted models is thus correlated with the evolutionary distance between the template (member of the family with known structure) and the target proteins (Tramontano et al., 2008).
Homology model techniques are the most reliable and accurate methods to generate structures (Meier and Söding, 2015 and Tramontano et al., 2008). Broadly speaking, comparative modeling can be considered as a cost-effective alternative tool when experimental structures are absent. Although homology modeling can boast a number of successes in many applications, it should be noted that the generated models are predictions and could present some inaccuracies.
In this subchapter we discuss the main advantages and limitations of the comparative modelling techniques (Busato and Giorgetti, 2016). It is well known that homology modeling methods depend strongly on both the sequence identity (SI) between the target and the templates and the accuracy of alignment. SI < 50% generally leads to structural divergence between the models and the actual experimental structure, measured as Cα atom RMSD, larger than 1 Å (Chothia and Lesk, 1986). Actually, two proteins with SI > 35% were shown to share the same fold (Orengo et al., 1997). Finally, with low target-template SI (between 10% and 25%), the comparative models might contain serious errors, thus, it is strongly recommended the introduction of experimental information such as ligand information, site-directed mutagenesis, and other experimental restraints to improve the accuracy of the model (Yarnitzky et al. 2010). These very general observations are often sustainable but can vary depending on the protein of interest. Indeed, it is well-established that some "superfolds" dominate the fold space. This implies that even very distant proteins, e.g. GPCR proteins, can share a similar 3D structure (Magner et al., 2015). The limitations/errors in homology modeling can be grouped into five categories (Fiser, 2010 and Palomba and Cavasotto, 2015):(1) Errors in side-chains modeling. These errors can introduce drastic changes in the side-chains involved in the ligand binding (Rodrigues et al., 2013). The latter highlights the importance of using existing ligand information in the homology modelling protocol (Yarnitzky et al., 2010). (2) Structural Deviations of a target region that is aligned correctly with the template. Even if the aligned segments of the model are correct; the target protein could present local structural differences than the template structure, indeed these divergences could be

due to artifacts in structure determination of the template in different environments, and not to errors in the alignment. Therefore, models might be improved by using unrestrained MD simulation in order to arrange the mis-folded regions (Schlegel et al., 2005) or by using multiple-template approaches (Srinivasan and Blundell, 1993). (3) Inaccuracies of target regions that do not have an equivalent segment in the template structures. Indels are an important issue in the homology modeling procedure. For instance, residues in incorrect positions or missing residue in the comparative models could be due to gapped residues in the core region. (4) Distortions or shifts of a region that is aligned incorrectly with the template structures. The quality of the alignment is one the major problems in homology modeling, especially when the SI falls below 20%. Misaligned regions correspond to errors in positioning the target residues on the template fold; resulting in an unreliable model. Multiple sequence alignment (MSA) and hidden Markov models (HMM) profiles approaches combined with manual inspection and curation of the alignments are strongly recommended to investigate possible errors and adjust key motifs in the alignment. (5) Templates are fragmentary or incorrect. The misfolded structure resulting from using an erroneous template or fragmentary template is the problem typically arises when SI is below 30%. Moreover, for large protein (>200 residues) the number of full-length templates is still sparse comparing with the many fragments of full-length proteins".

3.1.2 Steps in model generation

It is commonly accepted that the comparative modeling procedure involves the following steps (Fig. 3.1): (1) identification of proteins evolutionarily related to the target sequence, whose 3D structures are solved (templates); (2) target-template sequence alignment; (3) generation of an initial model for the target based on the template selected and the sequence alignment; (4) refinement of the model at the side chain, loops and backbone level by using MD simulations and incorporation of additional experimental structural information; (5) validation and evaluation of the model. These steps must be iterated to achieve an acceptable model.

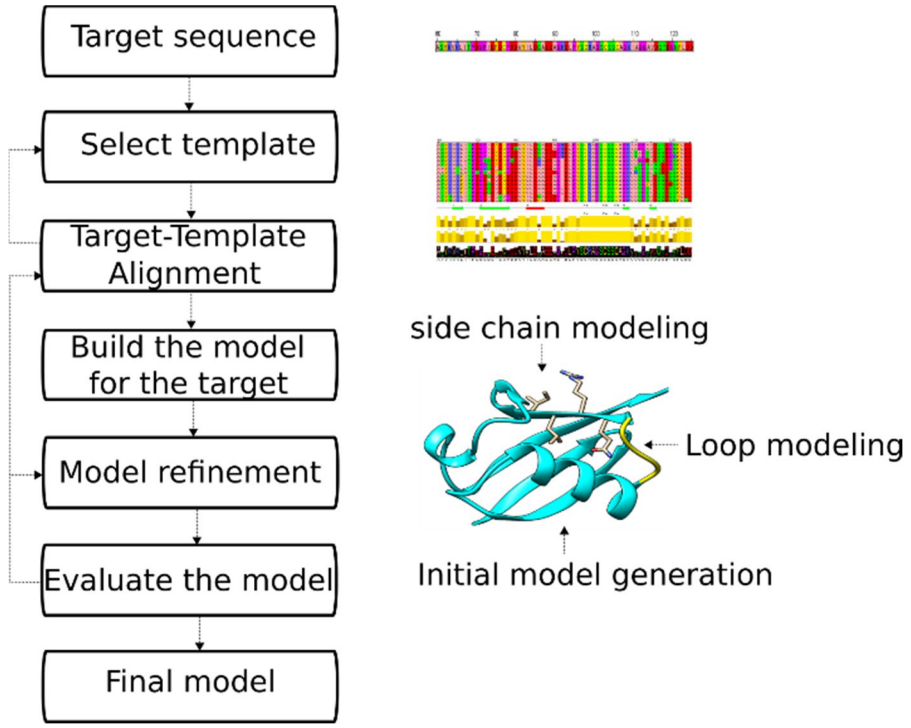


Figure 3.1. Workflow of the basic homology modeling procedure. Adapted from Busato and Giorgetti, 2016.

3.1.2.1 Template search and target-template sequence alignment

In homology modeling, template search and target-template sequence alignment should be considered as two independent and separate stages of the process. Actually, many homology modeling programs tends to integrate these two steps. This because the identification of a 3D structure used as template is based on sequence alignment aided by database search techniques. As indicated in the Eswar et al., 2008 paper, the sequence-structure relationship can be grouped into three different categories in the sequence similarity spectrum. The first group comprises the "easily detected relationships", characterized by SI higher than 30%. As shown in Sander and Schneider, 1991 paper, most protein pairs (90%) sharing more than 30% SI were found to have similar structure. The second group is the so-called "twilight zone" (Rost, 1999), which corresponds to relationships with statistically significant SI (10-30% sequence similarity). When the SI falls in the twilight zone, the statistical measure for the evolutionary relatedness of proteins can be unsure (Rost, 1999). The third group represents the "midnight zone" (Rost, 1999), corresponding to statistically irrelevant sequence similarity. Depending on different regimes, different sequence alignment methods can be used. Historically, the alignment methods can be divided in three group:

1) *Pairwise sequence alignment methods.* Pairwise sequence alignment methods search for the global or the best local alignments of two sequences (Eswar *et al.*, 2007).

While, the global alignments try to align the entire length of sequences, the local alignments only search highly similar portions within long sequences. By definition, pairwise alignments can only align two query sequences at a time, and they can be exploited to find, in a database, the sequences with high similarity to a query sequence. The two most famous pairwise sequence alignment methods aided by database search techniques are FASTA (Lipman and Pearson, 1985) and BLAST (Altschul *et al.*, 1997).
2) Profile-sequence alignment methods. With the profile-sequence alignment algorithms the query sequence is compared with a sequence profile, which typically represents a protein family and contains the evolutionarily relevant features of the family. The sequence profile is calculated from a MSA and it is represented as a position-specific scoring matrix (PSSM) (Henikoff and Henikoff, 1994 and Altschul *et al.*, 1997) or as a HMM (Krogh *et al.*, 1994 and Eddy, 1998). In this approach, instead of using only a single sequence, a frequency profile is used to filter the sequence information that is not conserved in the protein family. These algorithms are more efficient than pairwise sequence alignment methods to find all distant homologues. They can recover roughly twice the number of homologs under 40% of SI (Eswar et al., 2007). The BLAST variant i.e. PSI-BLAST uses this approach (Altschul *et al.*, 1997).
3) Profile-profile alignment methods. These methods work in two steps: i) they generate a sequence profile of the target and then they compare it to the sequence profiles of solved structures. As stated in the Remmert et al., 2012 paper, the profile-profile alignment algorithms are the most efficient approach to choose and align templates in the homology modeling process. This fact was confirmed by Prof. Marti-Renom, who pointed out that "Profile-profile methods detect about 28% more relationships at the super family level and improve the alignment accuracy by 15- 20%, compared to profile-sequence methods" (Marti-Renom *et al.*, 2004; Zhou and Zhou, 2005 and Söding, 2005). These algorithms are implemented in the main web servers and software for protein structure prediction. Indeed, HHpred, the first server that is based on the pairwise comparison of profile HMMs, turned out to be the best-scoring web server in template-based structure prediction (Mariani *et al.*, 2011) in CASP9.

3.1.2.2 Template Selection

Once a list of suitable structural homologues templates (retrieved from PDB database) and their alignments with target have been realized, at least five strategies exist to select the best template(s) (Tramontano et al., 2006). The main strategies are:

1) select the template that shares highest sequence similarity with the target.
2) build a "theoretical template" by using the average of the coordinates of all possible templates.
3) take the conformation of different regions from a given set of templates, in such a way that the local similarity is highest in each region;
4) create a structural prediction from each template and select the best model according to some criteria (see 3.1.2.4 Model Quality Assessment Section)
5) derive constraints from the templates and then create a structural prediction that satisfies these spatial restraints.

3.1.2.3 Model construction and refinement

Once a template-target alignment is made, the coordinates of the target can be built.
There are three methods to generate the initial structural model: i) Rigid body assembly; ii) Segment matching; iii) Satisfaction of spatial restraints.
Some popular computer programs for homology modeling (Martí-Renom et al., 2000) are based on rigid-body assembly method. In this strategy, a homology model of query portions is built up from a set of rigid bodies determined from the aligned 3D template. The most software based on this strategy are SwissModel (Biasini et al., 2014; Arnold et al., 2006; Kiefer et al., 2009 and Guex et al., 2009), RosettaCM (Song et al.,2013), WHAT IF (Krieger et al., 2009 and Vriend, 1990), Prime (http://www.schrodinger.com/Prime/), COMPOSER (Sutcliffe et al., 1987), CONGEN (Bruccoleri, 1993), ICM (http://www.molsoft.com/), PrISM (Yang and Honig, 1999) and 3D-JIGSAW (Bates et al., 2001).
SegMod (Levitt. 1992) is the most notorious software based on segment matching approach. The latter builds the crude model by using a subset of atomic positions e.g. Cα from the 3D template (Palomba and Cavasotto, 2015).
Finally, the most famous program, which generates models by satisfaction of spatial restraints is MODELLER (Fiser et al., 2000; Martí-Renom et al., 2000; Šali and Blundell, 1994 and Webb and Sali , 2014). MODELLER was used in this thesis, so we decided to provide a short description. The algorithm uses a set of spatial constraints such as interatomic distances, φ and ψ and backbone angles recovered from the experimentally solved structure and subsequently applies them to the target sequence, each with an associated probability density function (pdf 's) (see Fig.3.2). These structurally-derived constraints are so combined with general spatial constraints, derived from a CHARMM force field, into an objective function. Therefore, the model is built minimizing the violations of the geometric restraints. In MODELLER, the minimization step is carried out in two phases. The first uses the variable target function method. Then, the structural prediction is refined by MD with simulated annealing (SA).

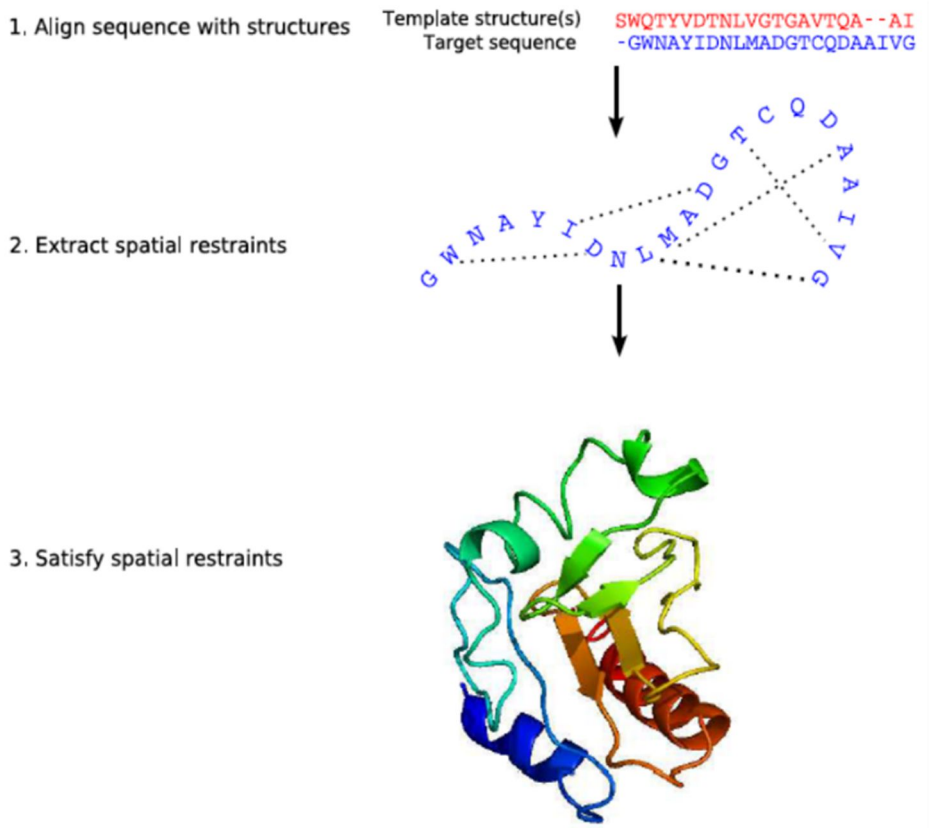


Figure 3.2. Basic steps of the MODELLER procedure. Adapted from the MODELLER 9v12 manual (https://salilab.org/modeller/9.15/manual.pdf).

3.1.2.4 Model quality assessment

The 3D structural predictions may contain errors, thus some features must be checked with attention in the model. The main model quality assessment factors/tools are briefly presented as follows:

- *Sequence identity.* The percentage of SI between target and template is a good predictor of the accuracy of a model. Model accuracy steadily increases with increasing SI.
- *Stereochemistry check.* The most common programs that evaluate the stereochemistry (e.g., bond-lengths, bond-angles, backbone torsion angles, and non-bonded contacts) of the structural prediction are PROCHECK (Laskowski *et al.*, 1993) and WHATCHECK (Hooft *et al.*, 1996). As said by Prof. Fiser, even though errors in stereochemistry are usually unusual and less informative than errors identified by statistical potentials, a cluster of stereochemical errors might indicate that there are larger errors (e.g., alignment errors) in that portion (Fiser, 2010).
- *Global model quality estimation*. Global indicators of the quality provide a pseudo-energy for the structural prediction. The most common global indicators are: 1) DFIRE (Zhou and Zhou, 2005), an all-atom distance-dependent statistical potential; 2) QMEAN (Benkert *et al.*, 2009) a composite scoring function for model quality estimation and 3) DOPE (Shen and Sali, 2006), a statistical scoring

function for model assessment based on an atomic distance-dependent statistical potential derived from ca. 1500 crystallographic structures.

- *Local model quality estimation*. Here, the accuracy of specific segments of the model is checked, identifying potentially incorrect portions in the structural prediction. The most famous tools are: 1) ProQres (Wallner and Elofsson, 2006), an artificial neural network trained to guess the local 3D prediction quality based on some structural characteristics; 2) ANOLEA (Melo and Feytmans, 1998), a statistical potential that can be exploited to investigate the packing quality of the model and 3) GROMOS (Christen *et al.*, 2005), an empirical force field that is main used to calculate the energy of each residue in the model.

3.2 Molecular Docking

3.2.1 Basic concepts of molecular docking

Proteins do not occur as isolated but are exposed to many interactions. In order to predict the biological function/activity of a protein, a detailed study of its interactions with its partners (proteins or small organic molecule) is required. The computational prediction of multimolecular complex structures is called *"docking"*. The binding pose of a ligand against the receptor can be uniquely described by its state variables: a number of parameters that fully describe the ligand position. The state variables of a ligand represent its coordinates, orientations, flexibility and structural conformations. A state variable represents one degree of freedom in a multidimensional search space, and their bounds describe the extent of the search.

Historically, simplified algorithms were developed to reduce the docking process, by keeping rigid the molecules along the docking protocol. However, it is important to highlight that these methodologies might produce unreliable results (poses) when the starting conformation of the ligands is wrong.

Broadly speaking, all the docking algorithms share two key components: i) a scoring function, which is maximized (minimized) to rank the complexes and ii) a search method to explore the state variables. In principle, the docking technique consists in a search of the conformational space of the most energetically favorable pose of the ligand. In practice, this would mean a full exploration of the components' energy landscape, resulting in an unrealistic approach due to the high computational time and cost. A classic example is the "virtual screening", a methodology that requires the calculation of protein binding poses of many ligands. Empirical, force field-based or knowledge-based scoring functions have been developed to reduce the size of the conformational space. These scoring functions can be divided into two groups: i) systematic and ii) stochastic.

Prof. Garrett M Morris in the "docking and virtual screening" book (Morris et. at., 1998) clearly defined these two terms as follows: while systematic search methods sample the search space at predefined intervals and are deterministic, the stochastic search methods iteratively make random changes to the state variables until a user-defined termination criterion is met.

The search methods can be further divided in "local" and "global" depending on how extensively they can explore the search space. While the local search methods attempt to detect the nearest or local minimum energy to the current conformation, the global search

methods look for the global minimum energy within the defined search space (Morris *et al.*, 1998).
In this thesis, state of the art docking programs have been extensively used, thus in the following sections each docking program is described.

3.2.2 Information-driven docking with HADDOCK

The docking procedure is based on the use of constraints resulting from experimental information in order to predetermine the surface of interaction between the macromolecules in question. HADDOCK (Van Zundert et.al, 2016 and Dominguez et.al., 2003) can be considered the most successful software that uses an information-driven approach, written and maintained by Alexandre Bonvin and coworkers. Originally conceived for protein-protein docking; now HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) has been applied at small molecule/protein docking. However, when small organic molecules are included in the docking phase, the force field topologies and parameters for these molecules must generated before to run a docking experiment. Two state of the art software i.e. PRODRG (SchuÈttelkopf and Van Aalten, 2004 and Van Aalten et.al., 1996) and ACPYPE (da Silva and Vranken, 2012) can perform this task.
While PRODRG (SchuÈttelkopf and Van Aalten, 2004 and Van Aalten et.al.,1996) is based on a database search method to parametrize the molecules using OPLS-like parameters (Paulechka et.al, 2012), ACPYPE (da Silva and Vranken, 2012) uses ANTECHAMBER (Wang et.al., 2006) with a semi-empirical quantum calculation method for the partial charges.

Ambiguous Interaction Restraints (AIRs). The experimental data (obtained by site-directed mutagenesis, NMR, bioinformatics predictions, etc.) are introduced in the form of ambiguous constraints of interaction (AIRs, Ambiguous Interaction Restraints), defined as ambiguous distances between all the residues involved in the interaction. An AIR causes a residue on the surface of a biomolecule that is lead in the vicinity of another residue or group of residues on the molecule partners, when they form the complex. An AIR is defined as an ambiguous intermolecular distance (di_{AB}) with a maximum value of 3 Å between any atom m of an active residue i of molecule A (m_{iA}) and any atom n of both active and passive residues k (N_{res} in total) of molecule B (n_{kB}) (and inversely for molecule B). The effective distance d_{iAB}^{eff} for each restraint is calculated using the equation:

$$d_{\mathrm{iAB}}^{\mathrm{eff}} = \left(\sum_{m_{i\mathrm{A}}=1}^{N_{\mathrm{atoms}}} \sum_{k=1}^{N_{\mathrm{res}}\mathrm{B}} \sum_{n_{k\mathrm{B}}=1}^{N_{\mathrm{atoms}}} \frac{1}{d_{m_{i\mathrm{A}} n_{k\mathrm{B}}}^{6}} \right)^{(-1/6)}$$

Equation 3.1. Effective distance (di_{AB}^{eff}) for each restraint in HADDOCK

The N_{atoms} indicates all atoms of a given residue and N_{res} the sum of active and passive residues for a given molecule. In this way, the passive residues do not have direct AIRs

to the partner molecule but can satisfy the partner molecule active restraints. A $1/r_6$ sum averaging is used because this mimics the attractive part of a Lennard-Jones potential and ensures that the AIRs are satisfied as soon as any two atoms of the two molecules are in contact. The 3Å limit represents a compromise between hydrogen-hydrogen and heavy atom-heavy atom minimum van der Waals distances. The use of AIR allows HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) to search through all the possible configurations around the interacting site defined by the information that we have about the problem in question. The AIRs are incorporated as a term of additional energy to the energy function (scoring function) that is minimized during the docking. It will still set an upper limit of distance, typically 2 Å, below which the constraint is satisfied and the attractive force is zero; if the limit is exceeded, the pairs of atoms are subjected to an attractive force that leads them to approach. Because some interatomic distances inversely contribute to the actual distance, an AIR is typically satisfied if a pair of atoms, part of a constraint, is between 3 and 5 Å, depending on the degree of ambiguity. In the definition of those residues, one can distinguish between "active" and "passive" residues. The "active" residues are essential for the interaction, based on experimental data. The "passive" residues are all solvent accessible surface, neighbours of active residue. These residues must be identified and specified by the operator. If experimental data are missing, bioinformatics is needed in the definition of active and passive residues. The most important bioinformatics tool for prediction of protein-protein interface residues is CPORT (de Vries and Bonvin, 2011). The latter combines six interface prediction methods WHISCY (de Vries et. al., 2006), PIER (Kufareva et.al., 2007), ProMate (Neuvirth et.al.,2004), cons-PPISP (Chen and Zhou,2005), SPPIDER (Porollo and Meller, 2007), and PINUP (Liang et.al., 2006) into a consensus predictor. CPORT (de Vries and Bonvin, 2011) results can be exploited as active and passive residues in HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003).

Docking Protocol. The program is limited to optimize the interaction between the two molecules in the space of the constraints provided by the operator. The protocol of docking, which requires the pdb files of the free molecules and AIRs, comprises three steps:

1. randomization of orientations and rigid body energy minimization (EM);
2. semi rigid simulated annealing in torsion angle space (TAD-SA);
3. final refinement in Cartesian space with explicit solvent.

The first docking step in HADDOCK (Van Zundert et al., 2016 and Dominguez et.al., 2003) is a rigid body energy minimization. First the molecules will be separated by a minimum of 25Å and rotated randomly around their centre of mass.
The rigid body minimization is performed stepwise: 1) four cycles of rotational minimization in which each molecule (molecule and associated solvent in case of solvated docking) is allowed to rotate in turn; 2) two cycles of rotational and translational rigid body minimization in which each molecule and associated solvent is treated as one rigid body.
The second docking step is semi-flexible SA. The best 200 structures after rigid body docking will be subjected to a semi-flexible SA in a torsion angle space. The semi-flexible annealing consists of several stages: 1) High temperature rigid body search; 2) rigid body SA; 3) semi-flexible SA with flexible side-chains at the interface; 4) semi-flexible SA with fully flexible interface (both backbone and side-chains). Practically, during the TAD-SA and the water refinement, the amino acids at the interface (side chains and

backbone) are allowed to move to optimize the interface packing. The interface amino acids allowed to move are defined by the active and passive amino acids used in the AIRs ± 2 sequential amino acids.
The third docking step is a flexible explicit solvent refinement. In this final step, the structure obtained after the semi-flexible SA is refined in an explicit solvent layer (8Å shell of TIP3P water molecules). In this step, no real significant structural changes are expected; however, the scoring of the various structures is improved. The final structures are analysed and each is assigned a score, called "HADDOCK Score", which is a sum of several energy parameters: term intermolecular electrostatic (Elec), a term of van der Waals (vdW), a term of desolvation (Dsolv), a term energy of AIR and a term of buried surface area (BSA).

The "rigid-body score" (RBS) is equal to:

RBS = 1 * Elec + 1 * vdW – 0,05 * BSA + 1 * Dsolv + 1 * AIR

Equation 3.2. Rigid-body score HADDOCK

The final score is:

FS = 1 * Elec + 1 * vdW + 1 * Dsolv + 1 * AIR

Equation 3.3. Final score HADDOCK

Subsequently, the results are grouped in clusters, based on the RMSD (Root Mean Square Deviation) of each structure with respect to the reference structure (the structure to lower HADDOCK Score), using a cut-off determined manually by the operator (generally, around 7Å).

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2}$$

Equation 3.4. RMSD HADDOCK

In equation 3.4, the δ is the distance between the pair of equivalent atoms N (generally, are used C_α and sometimes C, N, O, C_β). In essence, the RMSD is the quadratic differences of the average distance of all C_α of a molecule compared with the same equivalent C_α of another molecule and it measures the similarity of two 3D structures of two molecules following the superposition of the structures, whilst maintaining them rigid. To assess which model is best, one cannot rely on a single parameter, but some considerations must be made both on the energy, the frequency of a certain model (ie, how many times this model back in the dock, or how high the probability that is found in nature), and on the "visual appearance" of the complex generated. The HADDOCK (Van Zundert et al., 2016 and Dominguez et.al., 2003) program assigns to each structure and each cluster a score, HADDOCK score, which takes into account all the energy parameters. The complex is energetically stable as the HADDOCK score is lower.

Obviously, the more stable energy does not always indicate the goodness of the structure. This because it can be that the real model *in vivo* is not necessarily the more energetically stable. For this reason, in addition to considering the scores energy, the user should consider how many times a certain "type" of structure is returned. In other words, the cluster is created, grouping all the structures that have a similar RMSD, which results in a structural similarity. Depending on how a cluster is populated, we can understand how likely a certain type of interaction takes place. Taking into account the abundance of a cluster against his HADDOCK score can be a good method in understanding whether a structure is reliable. Of course, it is necessary to look at the pdb file of the various models generated to understand if the program has created models consistent with the input data. It can happen that, to minimize the energy, the software chooses not to interact with an extended area of a molecule: obviously, such a structure, despite having a very good HADDOCK score, is definitely not a good model.

3.2.3 Autodock suite

The Autodock suite is a state of the art package that is used in numerous laboratories and for different application such as docking and virtual screening. Indeed, as stated in the Forlì et.al., 2016 paper, Autodock can boast over 1,000 citations in the past year.
The Autodock suite, including source, is free available and comprises several tools e.g. Autodock (Morris et al., 2009), Autodock Vina (Vina) (Trott and Olson, 2010), Raccon2 (Forli et.al., 2016), AutoDockTools (ADT) (http://autodock.scripps.edu/resources/adt), AutoLigand (Harris et.al., 2008). In this thesis, two computational docking programs (Autodock and Vina) have been used and tested.
Vina (Trott and Olson, 2010) uses a simple scoring function and rapid gradient-optimization conformational search while Autodock (Morris et al., 2009) relies on an empirical free-energy force field and rapid Lamarckian genetic algorithm search method. Vina (Trott and Olson, 2010) is a very fast method but uses several approximations in the scoring function. Indeed, the latter does not include any electrostatic terms, hydrogen atoms and furthermore it uses spherically symmetric hydrogen bond potentials to model hydrogen bond interactions. Vina has shown to give good results when ligands have typical biological size and composition (Trott and Olson, 2010).
On the other hand, the Autodock score functions comprises physically based terms. Indeed, the latter includes explicit polar hydrogen, a term for the electrostatic contribution and a term for the directional hydrogen-bonding interactions. The Autodock score function is already optimized to give successful results for different biological systems (Forli et.al., 2016) and it is made available to the scientific community in the official web site (http://autodock.scripps.edu/resources/parameters) allowing the expert users to calibrate the scoring function parameters depending on the particular system under investigation (Forli et.al., 2016).
Both programs use stochastic search methods and predict a set of optimal docked poses, which are usually cluster spatially to verify the consistency of the results. However, it is worth mentioning that, by definition, stochastic methods do not guarantee that the global minimum is found.

Finally, to reduce the volume of the conformation space and thus the computational time, cost, both algorithms keep rigid the receptor molecule. As the matter of fact, the latter can be considered the most approximation in the Autodock/Vina based docking protocols.

Autodock docking steps. The Autodock docking (Morris et al., 2009) experiment includes at least seven steps, which are described as follows: 1) preparing the receptor molecule. In this step a .pdbqt file is generated, which contains information both hydrogen atoms and partial charges. By default Autodock uses "Gastaiger" method to calculate the partial charge. The Gastaiger" charge are calculated by using a fast empirical scheme. 2) preparing the ligand molecule. This step is similar to the previous step, but when the .pdbqt file is generated, the information of the ligand flexibility is included. 3) generating of the grid parameters file. Here the volume of the space (usually around the binding site of a receptor) that Autodock considers for the docking, is defined. This step generate the input file for "AutoGrid4" tool, which outputs several "map" and "grid" data file. 4) generating of the maps and grid data file. In this step, a set of maps and grid file are generated by AutoGrid4.2 script (Morris et al., 2009). 5) generating of the docking parameters file. Here, all the information for running Autodock are prepared. 6) running Autodock. 7) Results and analysis. Here, the complexes together with the estimate free energy of binding can be analyzed by using the graphical user interface ADT.

Vina docking steps: Vina (Trott and Olson, 2010) handles the same input file format i.e .pdbqt. Step1) and step2) remain unaltered, but Vina needs a particular file called "conf.txt" (see Fig.3.3), which contains all the docking parameters for the execution. Once the Vina job is launched, the program usually after few minutes returns the ligand poses (usually 10) with their estimate energy affinity.

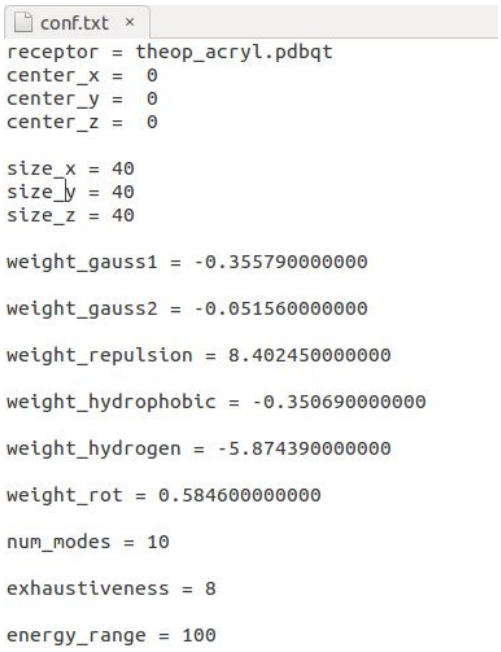


Figure 3.3. Screen shot showing a typical VINA configuration file. (For more details, see Vina tutorial (http://vina.scripps.edu/manual.html)).

3.2.4 LeapFrog (Sybyl)

LeapFrog algorithm (Tripos Inc., Tripos, St. Louis, MO, USA), a component of Sybyl (Tripos Inc) (v. 7.3) modelling suite, is used to screen virtual library of ligands/functional monomers (FMs) for their possible interactions with the receptor/templates. There are two different stages to calculate the energies: i) monomer-template complexation and ii) a system of scoring the complementarities between the monomer and template. In the latter, LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA) defines the template as the receptor binding site (using additional site-point matching scores, a system of scoring the receptor and ligand interactions).
This two-stage scoring is used to search a kind of regionality, therefore the "best fit" and the highest binding energy (in kcal/mol) will happen the best score is associated to complementary sites, which are nearby. The parameters in LeapFrog algorithm (Tripos Inc., Tripos, St. Louis, MO, USA) (see Fig.3.4) are standard and upon which the binding score is calculated. They have been optimised to favour hydrogen bonding capacity between small, finite molecules, and thus place less weight on other interactions such as hydrophobic, π–π or aromatic bonds.

Parameter name	Description	Default weight	Optimised weight
Join	Join different fragments	2	0
Fuse	Fuse candidate fragments	0	0
New	new ligand is started by aligning fragments	5	10
Fly	Alternative minimum energy ligand orientations	2	0
Twist	Conventional minimisation	2	5
Refine	Improves newly identified ligands	2	0
Bridge	Considers all fragments as bridges	2	0
Complement	Chooses moiety complementary to a cavity group as a ligand	2	0
Save	Saves ligands that satisfy specifications	2	5
Weed	Discards all but the 10 best ligands	0	1
Crossover	Generates best hybridizations among similar molecules	0	0
Prune	Can delete moieties of a known ligand on the basis of their energy with the receptor	0	0

Figure 3.4. Description of the relative move frequencies used by the Leapfrog algorithm to assign a docking mode and affinity score to the intermolecular complex.

3.3 Molecular Dynamics

In the previous sections, we focused only on static prediction techniques. In fact, we investigated how generate a 3D model of the protein from its sequence (homology modeling) and how generate a protein-protein complex or a protein-ligand complex (molecular docking). However, the biological processes occur in time. Thus, we need a method to predict the dynamical behavior of biomolecules. This method is called molecular dynamics (MD): a computer simulation approach where the time evolution of a set of interacting atoms is followed by integrating their equations of motion.

3.3.1 Choice of simulation approach: length and time scales

Molecular processes present time scale that span from the fs to the second and from pm to µm. In principle, we should simulate each biological system by using the Schrödinger equation. In this case, the computation time and cost would be infinite so, we must choose the correct computation approach in agreement with the time and lengths scale of the system that we want to investigate (see Fig.3.5). It is certain that we need to find a compromise between simulation speed and accuracy in the description of the system. Broadly speaking we can present three levels of detail:

1. Quantum mechanics (QM)
2. Atomistic molecular mechanics (MM)
3. Coarse-grained molecular mechanics (CG)

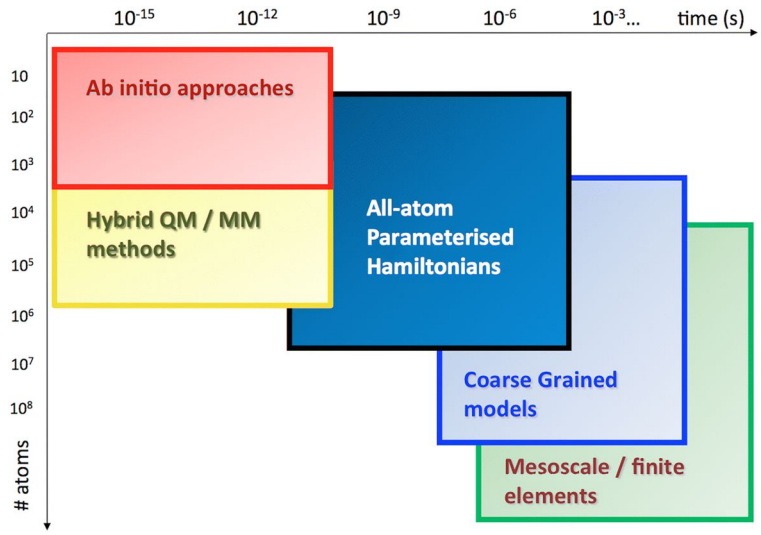


Figure 3.5. Accessible time- and size-scales for different computational methods in routine studies using state of the art implementations and architectures. Retrieved from Cascella and Vanni, 2015.

Originally, the methodology of MD was implemented before the 1960s and nowadays, many MD programs are available and the most famous are: GROMACS (Pronk et.al., 2013), AMBER (Salomon-Ferrer et.al., 2013), CHARMM (Brooks et.al., 2009) and

NAMD (Phillips et.al., 2005). GROMACS (Pronk et al., 2013) was used in this thesis; however, the following generic aspects described in the following sections are valid for all MD software[1].

3.3.2 Molecular dynamics: introduction

Molecular mechanics (MM) methods enable to study timescales spanning from 10^{-12} to 10^{-5} s, and systems of ca. 10^{6} atoms (Hug, 2013). Thus, systems e.g. proteins, in complex or not with ligand/protein in a solvent environment can be investigated, as well as the folding of short peptides. There are many MM techniques such as Monte Carlo (MC) simulations, energy minimization etc. However, all these approaches don't offer any information about the biomolecule behaviour over the time i.e. the dynamics.
MD can give information of the evolution of a molecular system in time. The smallest component of a system is the atom and these are considered as point in space, which follow the Newton's 2nd law: where the time evolution of a set of interacting atoms is followed by integrating (at discrete timestep) their equations of motion. The force field defines the force by which the atom interacts. The force field is a mathematical description of both bonded and non-bonded interactions, as depending on distances and angles between (two or more) atoms. This approximation implicitly includes all the electrons and their interactions but it does not provide any chemical reaction information. Again, MM is able to predict only conformational changes in the biological system, keeping the identity of molecules invariable along the simulation. In perfect conditions, the macroscopic (thermodynamic) calculable properties of the simulated ensemble can be reproduced by an extensive MD experiment. It does so under the ergodic hypothesis that if one allows the system to evolve on the limit of infinite time that system will eventually pass through all possible states.
Actually, the ergodicity is not achieved so, usually, it is a good compromise consider the convergence of a simulation when there are not alterations in the observables in time.

[1] In this section, we will provide information about the MM and CG approaches because they were extensively used in this thesis. Although we are aware that QM methods are very important in the field, they are out of the aims of this thesis and thus are not described here.

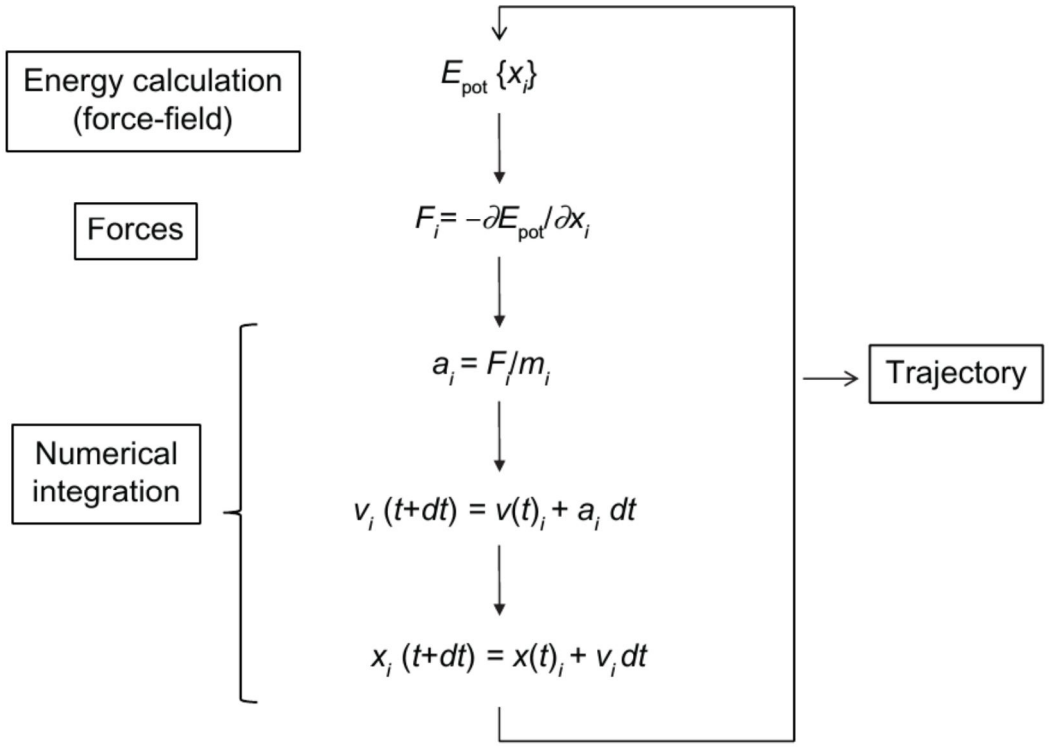


Figure 3.6. Workflow of molecular dynamics. The basic output is the trajectory, a file containing the positions (and possibly velocities and forces) of all the atoms in the system. Diagram retrieved from Sandal, 2016.

3.3.3 Periodic boundary conditions

The biological system under investigation is finite so boundaries must be considered. In principle, there are two options that can be adopted. In the first choice, the system can be inserted in a virtual box, thus it has a sort of "walls". Consequently, this could lead to surface effects (artifacts) in the MD simulation because the behavior of atoms in proximity of the walls is unphysical by definition. The solution to this problem is to use the periodic boundary conditions (PBC). On these terms, there are no rigid and impenetrable boundaries, indeed, the simulation occurs in a finite but unbounded topology, that corresponds to the 3-dimensional surface of a 4-torus. Atoms that get out in on one side, they get in on the opposite side, seamlessly. In practical terms, the box results to be infinitely replicated in each direction (see Fig.3.7). PBC can still produce minor artifacts, in fact, the user, during the MD setup, must carefully evaluate the volume of the virtual box to avoid interactions between the molecule of interest and its own specular images (that is something of unphysical). The box shape (cubes, parallelepipeds, rhombic dodecahedrons) should be able to minimize the holes in the simulation space. Thus, spherical PBC box are not possible. However, it is a good practice to simulate the system (usually in explicit solvent) in larger boxes than theoretical box size and moreover, the electrostatic treatment of PBC requires the box to be electrically neutral (Hug, 2013).

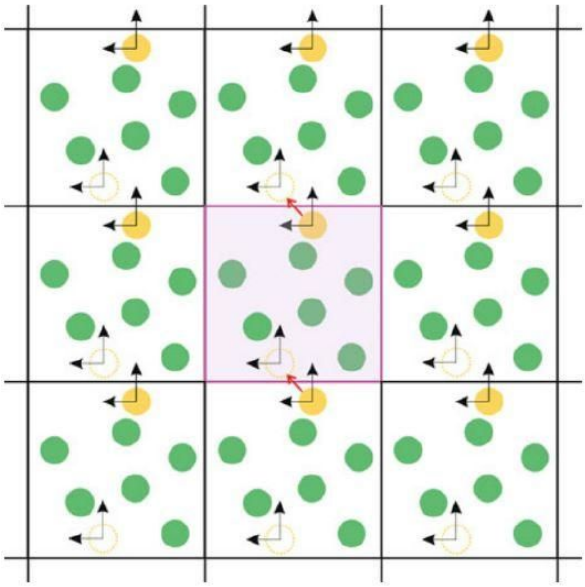


Figure 3.7. Schema describing the idea of PBC. The movement of the yellow particle (black arrows show the velocity components) is shown as moving from an edge of a box to the opposite edge. Retrieved from Hug, 2013.

3.3.4 Force Fields

A force field is a mathematical description of the force governing the interactions between atoms.
The interactions can be divided into two classes: i) bonded interactions and ii) non-bonded interactions.
Bonded interactions include the covalent bonds between atoms, and model bond stretching, rotations, bending, dihedral angles etc. Bonded interactions can involve 2,3 or 4 atoms and can be described for example by harmonic potentials.
On the other hand, non-bonded interactions include the electrostatic and van der Waals interactions, and they usually involve 2 atoms. Van der Waals interactions are described by Lennard-Jones or Morse potentials. The properties of an atom strongly rely on its chemical environment. In view of this point, a force field characterizes the parameters for different atom types even for the same chemical element (Leach, 2001). Given a force field, all the information of the system such as atom types, atom connectivity are defined as “topology” of the system.
The most famous force fields are: AMBER (Salomon-Ferrer et.al.,2013), GROMOS (Oostenbrink et.al., 2004), CHARMM (Brooks et.al., 2009) and OPLS (Jorgensen et.al., 1996).
GROMOS force field was used in this thesis and it is a united atom force field, this means that non-polar hydrogens are not modelled explicitly. Force fields' parameters can be calculated by using ab initio methods or experimental data.
While GROMOS is parametrized with the purpose of reproduce the free enthalpies of hydration and solvation (Oostenbrink et.al.,2004), AMBER (Salomon-Ferrer et.al., 2013) values are generated by ab initio QM calculations (Cornell et.al., 1995).

It is worth mentioning that since different force fields are generated by using different strategies, they cannot mixed in the MD setting up.

3.3.4.1 Coarse-grained methods

Many approaches have been proposed in order to save computational time and cost for MD simulations. One of these methods is called “coarse-grained (CG)”. The latter by coarse graining the molecular topology, reduces the multiple real atoms to a single coarse-grained component by a mapping process (see Fig.3.8) (Cascella and Vanni, 2015 and Shelley et al., 2001). In CG simulations, most of the MM basis and algorithms are valid. Indeed, GROMACS (Pronk et.al., 2013), the software for MD simulations, can work both with MM and CG atom representation; and the difference between MM and CG is in the force field used.

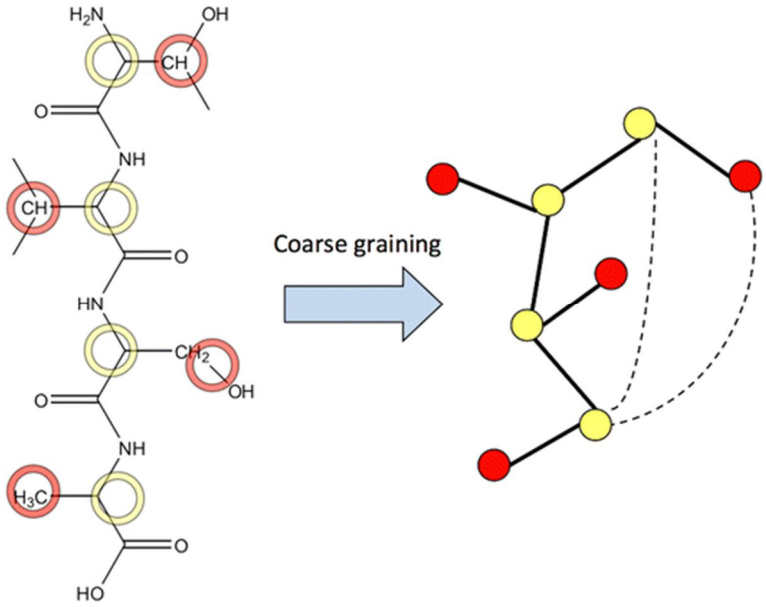


Figure 3.8. Example of mapping between an atomistic polypeptide chain (left) and a CG representation (right). Retrieved from Cascella and Vanni, 2015.

CG force fields and mappings are often generated using the same basis (reproducing experimental data sets, chemical-physical quantities), however different mappings and force fields (with different level of details) can be adopted based on what we want study (Cascella and Vanni, 2015 and Shelley et al., 2001). The two CG methodologies that have shown successful results are: MARTINI and Go models. In this thesis, MARTINI force field has been used and we will provide a brief description below.

3.3.4.1.1 MARTINI

MARTINI is both a CG force field and mapping protocol that has been exploited mostly for protein-lipid membrane MD simulation (Monticelli et.al., 2008 and Ceccon et al., 2015). The MARTINI parametrization follows the same idea of a MM force field and it is similar to GROMOS (Christen et.al., 2005) force field.

Non bonded force field parameters are described by the Lennard-Jones potentials. The latter attempt to reproduce the general chemico-physical quantities (called, oil/water partition coefficients) rather than a particular structural conformation.
On the other hand, bonded parameters rely on statistical polypeptide properties and are retrieved from PDB database. Generally, the mapping strategy is 4-to-1, i.e. 4 MM atoms correspond to 1 CG atom and the electrostatic is considered explicitly (Monticelli et.al., 2008). MARTINI includes the biomolecule structure information, since the secondary structure has to be set up at the start and restrained.

3.3.5 Solvation

Biomolecule should be model considering the solvent effect. The solvent in MD simulation can be considered in two ways:
i) Explicitly Solvent. Explicit molecules of solvent (usually water) are included in the simulation. This means that more atoms must be simulated, augmenting the computational time. Actually, the solvent models (e.g. SPC/E (Berweger et.al., 1995), TIP3P (Berendsen et.al., 1987)) still are not able to fully reproduce the water chemical-physical parameters. Indeed, in general water models show faster diffusion and lower internal structure than real water (Mark and Nilsson, 2001). Usually the water models with more than three atom points are used because they are able to better approximate the charge distribution of water (Mark, P. and Nilsson, L., 2001).
ii) Implicitly Solvent. In order to reduce the computational time and cost but at the same time keeping a certain level of accuracy in MD simulations, a new approach for solvent modeling has been developed. Implicit solvent methods model the solvent by a continuum potential and they usually treat the polar part as a continuous dielectric. The nonpolar component of solvation is estimated from the solvent-accessible surface (Chen et.al., 2008).

3.3.6 Cut-off

The calculation of non-bonded terms in the potential energy functions represents the most consuming part of MD simulations. The values of these terms grow in a quadratic manner with the number of atoms present in the system. For instance, in a system composed by 100 atoms, the number of interactions to be considered are about 5000 but if we have 10000 atoms, the interaction became ca.$50x10^6$. In these cases, to make more efficient the calculation, the interactions between atoms that are at a distance greater than a certain threshold (cut-off) can be neglected. Of course, ignoring completely the interactions that take place at distance greater (>cut-off) is a simplistic approach and can generate artifacts in the simulation. To overcome the problem several approaches have been developed. The most common approaches are three: i) truncation method, ii) shift cut-off method and iii) switch cut off method.
In the truncation method for the interatomic distances greater than the cut-off, the interactions are simply set to zero. The shift cut-off approach modifies the entire potential energy surface such that at the cut-off distance the interaction potential is zero. The switch cut-off approach reduces the interaction potential over a predefined range of distances. The potential takes its usual value up to the first cut-off and then it is switched to zero between the first and last cut-off

(http://www.ch.embnet.org/MD_tutorial/pages/MD.Part3.html).

3.3.7 Integration of the equations of motion

Given a system of N particles, and a force field:

$$U(\vec{r}_i), i = 1, \ldots, N$$

Equation 3.5. Force Field.

we have a 3N second-order differential equation set which represent Newton's equations of motion:

$$\vec{F}_i = m_i \ddot{\vec{r}}_i$$

Equation 3.6. Newton's second law.

In MD, we discretize differential equations as difference equations, integrated over discrete time steps. Anyhow, our goal is to simulate something where the time should be continuous. Thus, an important factor in MD simulation is the timestep. A large timestep would coarse grain the integration too much and would produce unphysical artefacts. As the matter of fact, if we decided to simulate the system with a small timestep, the computation time will be high. Therefore, the choice usually falls on the largest possible timestep that falls below half the characteristic time of the fastest motions of the system. The fastest movements are the vibrational movements (bond stretching and bending), which restrict our time step to 1-4 fs.
The integrator algorithm, in agreement to Newton's equations, should be also in theory time-reversible and symplectic (Hug, 2013 and Frenkel and Smit, 2002). An integrator algorithm is defined "symplecticness", when the phase space volume and total energy are preserved in time and so, it must follow the Liouville's theorem. The most famous symplectic algorithms are: Verlet (Verlet, 1967), Velocity Verlet (Verlet, 1968 and Swope et al., 1982) and Leapfrog (Hockney and Eastwood, 1988).

3.3.8 Thermostats

Given that Newton's equations conserve energy, the integration of Newton's equations of motions would direct the simulation to belong to a microcanonical (NVE) ensemble (Frenkel and Smit, 2002). However, we need to recreate situations that are close to the biological reality, where the temperature is preserved (NVT or NPT ensembles). To perform this, we need to mimic the coupling of our system to a heat bath. The latter is carried out by a thermostat algorithm, which corrects the velocities of particles in order to recreate, with different degrees of approximation, a canonical ensemble.

The simplest method would consist rescaling the velocities so that they conserve, at each step, the Boltzmann distribution at the fixed temperature. Velocity rescaling, however, would damp or erase the thermal fluctuations of the system, producing an artificial ensemble.[2]

[2] Many other algorithms exist, anyway, a detailed description of these is out of the scope of the present thesis.

Chapter 4

Computational Design and Optimization of Molecularly Imprinted Polymers

4.1 Introduction on Molecularly imprinted polymers (MIPs)

The common bioanalytical techniques concern antibodies, enzymes and receptors for the identification of molecules within complex matrices. However, the use of these strategies involves different problems: molecules are unstable, the available binding sites are very low, the affinity is difficult to regenerate after the molecular recognition and at last, it's difficult to produce specificity for analytes or prevent access to binding sites of other substances. Molecular imprinting technology is a technique used for the preparation of polymers that have specific molecular recognition properties for a given compound, its analogues or for a single enantiomer. In 1931, Polyakov investigated the polymerisation process of sodium silicate in water after the addition of ammonium carbonate (Alexander et.al., 2006). The authors reported the effects of different solvents (benzene, toluene and xylene) on the silica pore structure during the drying of a newly prepared silica. When H_2SO_4 was used as the polymerization initiator (acidifying agent), a positive correlation was found between surface areas and the molecular weights of the respective solvents. The silica particles produced were able to rebind this additive more than its structural analogues, thus exhibiting a sort of "memory" effect. In 1972 Wulff and Sarhan (Wulff and Sarhan, 1972) applied the same approach to organic polymers. In 1977 Wulff and his colleagues reported the first covalent imprinting. They synthesized 2:1 covalent conjugate of p-vinylbenzeneboronic acid with 4-nitrophenyl-a-D-mannopyranoside (the template), and copolymerized this conjugate with methyl methacrylate and ethylene dimethacrylate (a crosslinking monomer). After the polymerization, the polymer was broken and the 4-nitrophenyl-a-mannopyranoside was removed. The resulting polymer was able to bind selectively the aforementioned sugar. The complex conjugate, imprisoned in the polymer, had produced a template capable of selectively recognizing the molecule templating agent. Mosbach in 1981(Arshady and Mosbach, 1981) demonstrated that the covalent bond between the functional group and the templating agent was not strictly necessary in the process of polymer synthesis: non-covalent interactions were enough. In this way, by mixing the reagents in the reaction mixture, their non-covalent adducts were spontaneously formed. Indeed, the polymerization took place with the same satisfactory results, faster and more easily. Finally, Whitcombe at all. (1995) (Whitcombe et al., 1995) combined the advantages of the covalent bond (its extreme precision) with non-covalent bond method (the speed in forming and break the bonds with the mold) for the synthesis of polymers for cholesterol. The technique of molecular imprinting allows the preparation of synthetic polymers with specific binding sites for a target molecule. This can be achieved if the target is present during the polymerization process, thus acting as a molecular template. Monomers carrying certain functional groups are arranged around the template through either non-covalent or covalent interactions. Following polymerization with a high degree of cross-linking, the functional groups become fixed

in defined positions by the polymer network. Subsequent removal of the template by solvent extraction or chemical cleavage leaves cavities that are complementary to the template in terms of size, shape and arrangement of the functional groups. These highly specific receptor sites are capable of rebinding the target molecule with high specificity, sometimes comparable to that of antibodies. MIPs have therefore been named "antibody mimics". It has been shown that they can be substituted for biological receptors in certain formats of immunoassays and biosensors. They are characterized by high stability. In contrast to biomolecules, MIPs are usually stable at low and high pHs, pressures and temperatures (< 150 °C); in addition, MIPs are able to retain their recognition ability for longer periods of time than natural antibodies or receptors. Moreover, they are less expensive than biomolecules and easier to obtain, and they can be used in organic solvents. Finally, they can be synthesised for diverse classes of substances, such as ions, proteins, drugs and even yeast cells and erythrocytes. MIPs have some limitations connected with the methods of their production and the final format of the polymer: they present a better performance in organic medium than in the aqueous medium and they don't have a standard procedure for their preparation (Sellergren, 1997).

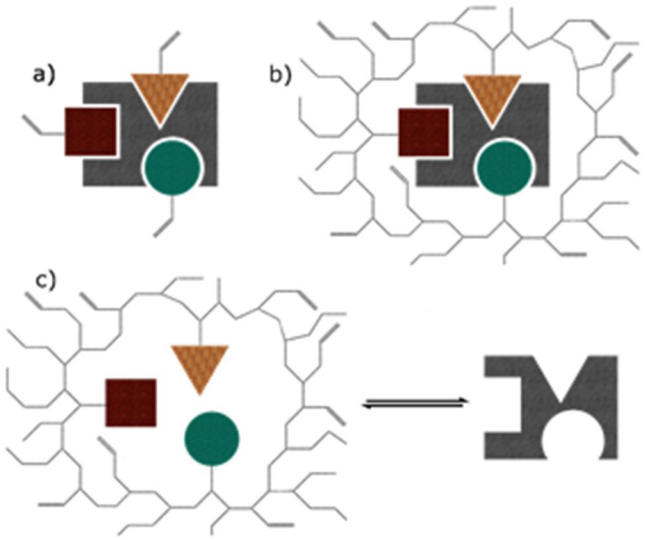


Figure 4.1. Scheme of molecular imprinting. MIP synthesis: a) complex formation between functionalized monomers and template; b) polymerization; c) MIP with template/target. Adapted from Cowen et al., 2016.

Representation of the general scheme of non-covalent molecular imprinting is shown in Figure 4.1. For a template molecule (or target or print molecule), appropriate FMs are chosen and allowed to form a self-assembly construct. By copolymerization with a cross-linking monomer, a polymer network is formed in which the self-assembly is set. Thereby, the position and the spatial conformation of the monomers are constructed according to the template. Then, the embedded template can be extracted and rebound to the molecularly imprinted polymer (MIP).

4.1.2 General applications of MIPs

Although the first studies of MIP technology began in 1970, the progresses made in this field are recent: in fact, the poor performance of MIPs in aqueous environments and the absence of any general procedures for MIPs design created some difficulty in the

development of this technology. An important area in which MIPs could be applied is separation: affinity separation, solid-phase extraction (SPE), chromatography and separation under extreme conditions. MIPs, indeed, could be a solution to the removal of harmful substances such as pesticides, endocrine-disrupting compounds and heavy metals from both waste and drinking water. Due to their selectivity can be used as adsorbents in the techniques of SPE. The extraction by adsorption is a physical process between a solid phase and a liquid phase in which the solid phase has a greater affinity for the compound to be isolated with respect to the solvent, in which the same compound is dissolved. When the sample passes through the solid phase, analytes are concentrated on the surface of the adsorbent material while other compounds, present in the sample, elute without interacting. The result is the purification and concentration of the substances isolated from the adsorbent material. This can be achieved through specific interactions between the functional groups of the compounds and the substrate of the solid phase. The MIPs have binding sites embedded in a solid matrix that resist to very difficult environmental conditions (high temperatures and pressures, non-aqueous solutions and pH extremes). They can be regenerated and reused without reducing their capacity even after prolonged use. Another important advantage is the ability to analyse a large number of samples due to the high density of binding sites, higher than antibodies used as immunoadsorbent. The possibility to obtain a greater recovery of analytes, to lower the limits of detection, the high stability, the greater selectivity, the possibility of working with organic solvents and the low production cost, make this approach very promising compared to current protocols for SPE. The second application of MIPs is sensor design. The high stability of MIPs renders them as prime candidates to replace biological receptors or they can be specially synthesised for analytes for which there is no available receptor or enzyme, or as a cheaper alternative to existing ones. Another application of MIPs is catalysis: typically, the most recognized area for their application is the design of zeolites (molecular sieve) implicated in the refining and processing of gas and oil. MIPs technology has the underlying potential for drug release matrices and the modification of the surface biocompatibility of medical devices. However in order to clearly establish the commercial prospects of the technology a greater depth of biomedical information is required (Makoto et.al., 2004).

4.1.3 Current methods of MIPs production

4.1.3.1 Bulk Polymerization

To suit the ultimate application desired, MIPs can be prepared in a variety or physical forms. Conventionally, the method for preparing MIPs is through solution polymerisation and it is followed by the mechanical grinding of the resulting bulk polymer that has been generated into small particles. This is followed by the sieving of the particles into various and desired sizes, with range from about 45 and 100 μm in diameter. In principle, this method is fast, simple and it does not require neither sophisticated equipment nor particular operator skills, thus, it is seen to be, by far, the most popular method (http://digilander.libero.it/Imprinted_Polymers/MIPs/sintesi_bulk.htm).
In opposition, though bulk polymerisation is simple and the optimisation of imprinting circumstances are moderately straight forward, there are still many shortcomings present in relation to this protocol. A negative impact on chromatographic performance and lower

MIP loading capacity are the results of the last sieving step. This step could leave the obtained particles in highly irregular sizes and shapes (Yan and Row, 2006).

4.1.3.2 Precipitation Polymerization

The method of polymerization by precipitation is very similar to that of the synthesis in bulk and the only real difference is the different relationship between the amount of porogen solvent (in this case much greater) and other compounds present in the reaction mixture. This excess of solvent porogen prevents the formation of a solid block. This technique is based on the same principle as the synthesis in bulk, but it has the advantage of leading directly to the production of polymeric microspheres of regular size and shape (http://digilander.libero.it/Imprinted_Polymers/MIPs/sintesi_precipitazione.htm).
This avoids the process of breaking, crushing and sieving necessary to obtain a stationary phase desired size. The microspheres have dimensions of the order of 1μm and they are especially useful as a stationary phase for HPLC analysis (Yan and Row, 2006).

4.1.3.3 Grafting polymerization on monodisperse polymer beads

The main problem during the MIPs preparation with the technique in bulk is the production of a irregular particulate mesoporous not suitable for the production of stationary phases for HPLC analysis. The grafting technique produces stationary phases with high selectivity that are useful for both, chromatography and solid phase extraction (SPE). A thin layer of polymer is created around preformed polystyrene spherical beads, with controlled porosity. In this way, good levels of selectivity and binding capacity are obtained. The MIPs preparation process involves at first the synthesis of beads, also produced by polymerization.
When the beads are ready, they can be covered with a thin layer of the desired polymer (http://digilander.libero.it/Imprinted_Polymers/MIPs/sintesi_grafting.htm). To do this, the beads are reacted and covered with a photosensitive radical initiator, the iniferter (sodium diethyldithiocarbamate trihydrate). Once ready, the beads are reacted with the polymerization mixture for obtaining the final MIP product (Yan and Row, 2006).

4.1.3.4 Emulsion synthesis

Emulsion synthesis technique demonstrates that stationary phases with high selectivity are useful for both, chromatography and the solid phase extraction (SPE) and, unlike the grafting technique, it is possible to directly prepare the polymer without having to first produce the beads (Yan and Row, 2006) (http://digilander.libero.it/Imprinted_Polymers/MIPs/sintesi_emulsione.htm).

4.1.4 Rational Design of MIPs

4.1.4.1 Combinatorial approach

Several variables such as complementary interactions between the template and the functional/crosslinking monomers, stoichiometry and concentration of the template and monomers, which solvent is used and the temperature of the imprinting process may affect the MIP's selectivity and capability. Therefore, a preferable way for defining the optimal MIP formulation is the combinatorial approach. For instance, with Takeuchi et al, (Takeuchi et.al., 1999) by a recently developed semi-automatic system, combinatorial libraries of MIPs were prepared and screened for high affinity and selectivity to the original template. MIPs as synthetic receptors for molecules e.g. triazine herbicides, ametryn and atrazine, were equipped by the combinatorial molecular imprinting using an assorted amount of two FMs such as methacrylic acid (MAA) and 2-(trifluoromethyl)acrylic acid (TFMAA). It can be seen, through the examination of the MIP libraries, that the imprinting efficiency is different for each triazine herbicide depending upon the FM used. For example, MAA is ideal for the atrazine receptor preparation and TFMAA for the ametryn receptor preparation (Takeuchi et.al., 1999). It can be seen, through the results, that the proposed high-throughput combinatorial molecularly imprinting method is a promising system for discovering ideal conditions of MIP preparation for specific molecules (Lanza and Sellergren, 1999).

4.1.4.2 Thermodynamic approach

The quality and performance of the polymer product is determined by the nature of the template and the monomers and by the polymerisation reaction itself. Furthermore, the direct function of the mechanisms and extent of the monomer–template interactions present in the pre-polymerisation mixture determine the quantity and quality of the MIP recognition sites. Paradigms developed by Prof. Jenks have been of fundamental significance to the methodologies used by a number of researchers in determinations to identify these physical factors. Semi-empirical approaches have been independently developed by groups such as that of Andrews et al. on the estimation of ligand–receptor binding constants and on the factorisation of energetic contributions to binding (Andrews et.al., 1984).
Williams et al. development (Williams et.al., 1991) of general thermodynamic treatment can be employed as a foundation for improved understanding of the recognition events involved in MIP synthesis also to ligand–MIP binding events. Thermodynamic considerations of MIP recognition were also studied by Prof. Nicholls (Nicholls, 1998). The physical terms that rule a binding event were attempted to be defined in these approaches. Prof. Nicholls explained that the extent of template complexation at equilibrium is governed by the change in Gibbs free energy of formation of each mode of template–FM interaction. The monomer(s)–template complex is not exposed to conformational strains and unfavourable van der Waals interactions, as the prearrangement phase is under thermodynamic control. Additionally, because of its high degree of cross-linking, MIPs only undergo partial deviations in its conformation during the recognition of the template. Hydrophobic interactions can be considered negligible due to the information both polymerisation and rebinding processes occurred generally in

lipophilic solvents. Prof. Nicholls recommended a streamlined version of the energetic contributions to ligand–receptor interactions:

$$\Delta G_{\text{bind}} = \Delta G_{\text{t + r}} + \Delta G_{\text{r}} + \Delta G_{\text{vib}} + \Sigma \Delta G_{\text{p}}$$

Equation 4.1. Gibbs Free Energy

The Gibbs free energy changes are: ΔG_{bind}; complex formation; ΔG_{t+r}, translational and rotational; ΔG_r, restriction of rotors upon complexation; ΔG_{vib}, residual soft vibrational modes and $\Sigma \Delta G_p$, the sum of interacting polar group contributions (Karim et.al., 2005).

4.1.4.3 Computational Design using Sybyl protocol

The computational methods for rational design of MIPs, developed by the Polymer Group led by Professor Piletsky at Cranfield, UK (now at Leicester), consist of four steps, each of which can be used autonomously. The first step consists of the formation and modelling of the template molecule. The first step consists of the formation and modelling of the template molecule. Primarily, a molecular model of the template model is built, at that time the charge for every atom is calculated. Then, the structure of the template is refined using MM. The second step consists of the construction of a virtual library. Enclosed within this virtual library, there are molecular models of FMs possessing both polymerizable residues and residues. Theses residues are able to interact with the template through electrostatic, hydrophobic van der Waals forces, dipole-dipole interactions or reversible covalent bonds. There are many polymerizable compounds (about 4000) which can be used to build a large monomer database but many of them have same chemical properties, therefore, it is more suitable to create a small monomer database for computational reasons. The third step is the screening of the virtual library. This screening is done by means of the LeapFrog algorithm (Tripos Inc., Tripos, St. Louis, MO, USA). Each of the monomers in the database is evaluated to discover the interactions with the template. The monomers giving the highest binding energy and capable of forming the strongest complexes are selected and used in the fourth step to investigate their interaction with the template. In the fourth step, also called the refining step, many copies of the monomer chosen in the prior step, are positioned around the target in a virtual box. Here, a minimisation (MM) or SA (MD) process can be completed to investigate the monomer-template interactions. When the experiment is finished, the number and position of the monomers are analyzed; and the type and quantity of the monomers participating in the monomer-template complex then determines the ratio of template and monomers to use for the polymer preparation (Chianella et.al., 2002 and Subrahmanyam et.al., 2012).

4.2 Aims of the research described in this Chapter

- to design protocols for the predictive rational design of MIPs for practical applications:

  i) development of a publicly accessible web server interface aimed at the intelligent design of MIPs (section 4.3). Several application cases are offered as well (section 4.3.2).

  ii) implement a new algorithm which mimics a radical polymerization mechanism for application in the design of MIPs (section 4.4)

4.3 Development of an innovative web server aimed at the rational design of MIPs

4.3.1 Workflow

Our protocols use state of the art, proven bioinformatic tools (see Fig.4.2) in an easy user interface.

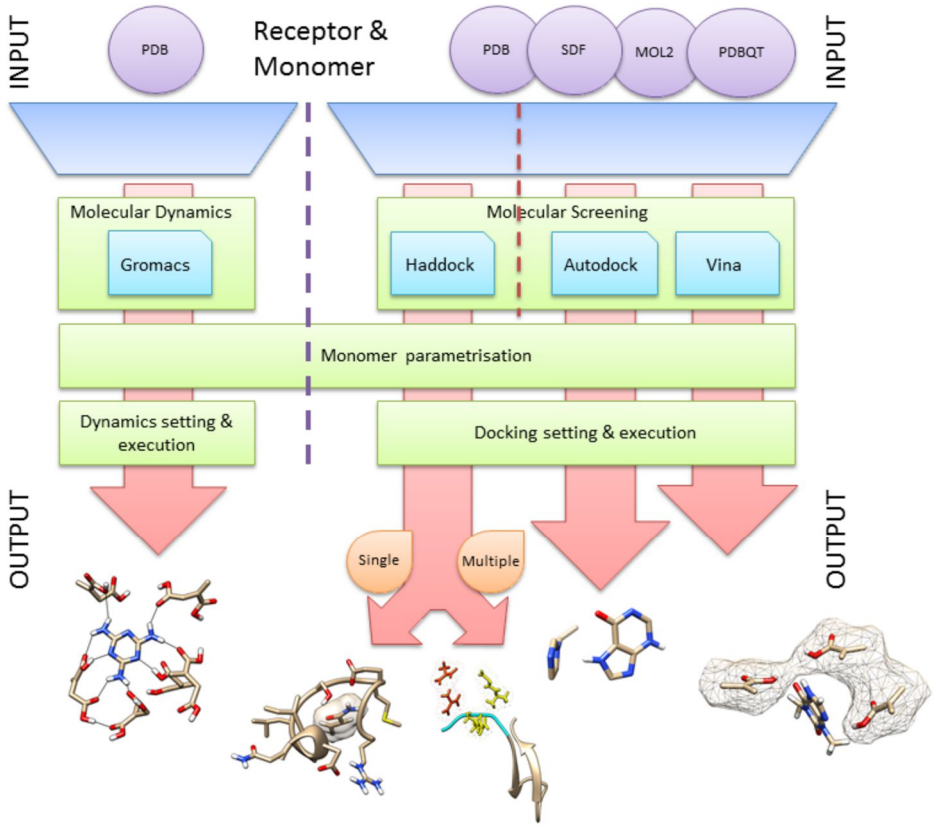


Figure 4.2. Intuitive representation of MIRATE web server. State of the art software such as HADDOCK (van Zundert et al., 2016 and Dominguez et al., 2003), Autodock (Morris et al.,

2009), Vina (Trott and Olson, 2010) and GROMACS (Pronk et al., 2013) are used in the MIRATE web server.

Given the most used FMs and the target molecule, our server suggests the putatively most affine monomer and the optimal stoichiometric ratio. Within the server the MIPs' design consists of three steps: 1) parametrization of the input molecules (receptor and monomers) with state of the art approaches, 2) FMs selection and screening of "virtually imprinted receptors" for rebinding of the molecular template by optimization of three established docking protocols 3) refinement step to calculate the ideal monomer/template stoichiometric ratio by using a novel molecular mechanics method. The web server includes a virtual library containing the most commonly used parametrized FMs, in ready-to-dock, and 3D formats. The web server allows a strong user intervention at different stages offering, to the user, the choice between corrected parameters for MIPs modeling and default parameters. In step 2), the small organic molecules screened against a template or a virtually imprinted receptor are ordered by binding energy/score. Furthermore, for each molecule, the web server provides both the complexes and the docking cluster information. This information is fundamental for wet-lab candidate preparation. In step 3), the server provides information about the main monomers-template interactions (hydrogen bonds) and the number of monomers interacting with template. Finally, the type and quantity of the monomers participating in the complex determine the ratio of template and monomers for the polymer preparation.
The MIRATE web server pipeline is briefly resumed in Fig.4.3 (for a detailed illustration of every single web-server component see Fig. S4.1.a and Fig. S4.1.b).

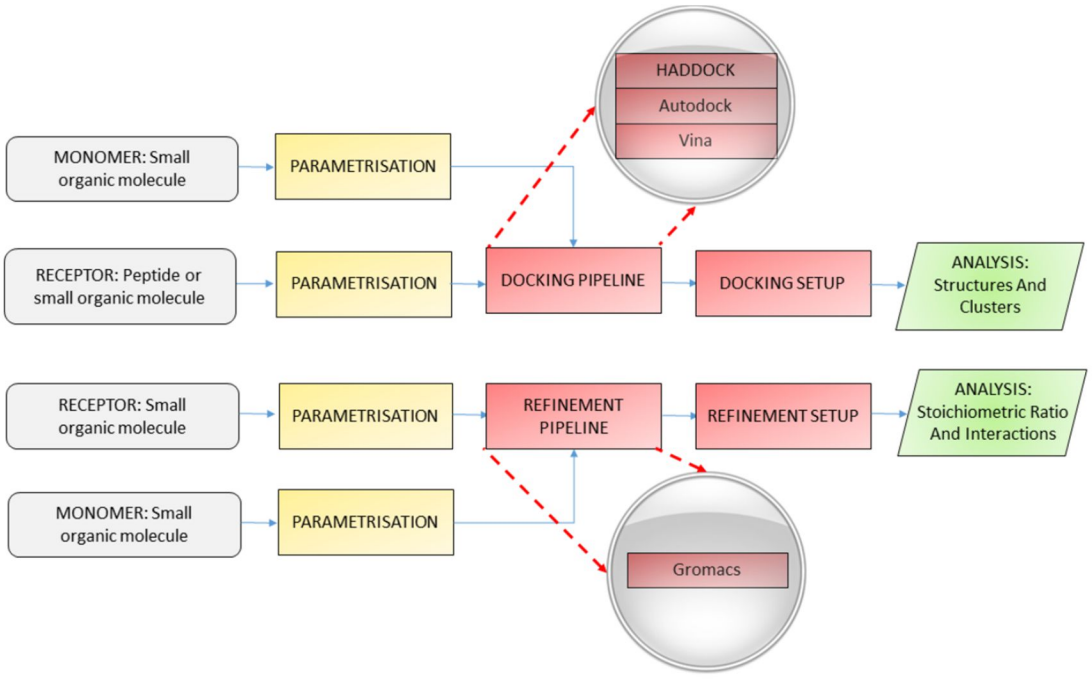


Figure 4.3. Flowchart showing the MIRATE Web server pipeline.

The MIRATE Architecture. MIRATE provides a whole series of facilities to perform molecular (i) docking and (ii) refinement, supporting both peptides and small organic

molecules. Since MIRATE is a web-oriented platform, it does not require any specific library or adjustment to be used, thus allowing users to focus only on experiment setting and execution. MIRATE *user interface* (UI) relies on latest web technologies, such as HTML5, JQuery (Javascript Framework) and Bootstrap. The *user interface* is fully responsive and interactive, adapting its content (data) to be visualized through any device, even mobile devices. The whole MIRATE platform is organized into two main blocks: The MIRATE web server designed to collect all submitted user jobs, and a computational cluster designed for intense calculus.
Submitted jobs are collected in a special queue, and later send to the cluster. Cluster is equipped with TORQUE Resource Manager, used to distribute and control the batch jobs. In such a way, the MIRATE web server is relieved from intensive computational tasks, providing a reactive system.

4.3.1.1 Input preparation

Receptor preparation. Users can upload a molecule as a template, which could be either a small organic molecule or a "virtual imprinted receptor" (Bates et al., 2016) or a peptide.

- **i) HADDOCK-based approach.** In the HADDOCK pipeline, the peptide structure with standard aminoacids, free of ions, ligands, and up to one chain, must be uploaded in .pdb format and its integrity is checked before proceeding to the docking step (Fig.S4.2). HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) is a free state of the art program designed for treat protein/peptide.

ii) AutoDock/Vina-based approach. When a small organic molecule or "virtual imprinted receptor" or peptide is uploaded as a template, the computational screening is performed by using Autodock (Morris et al., 2009) or Vina (Trott and Olson, 2010). These two programs are included in the Autodock suite and are state of the art software mainly designed to work with small organic molecules (Forlì et al., 2016). The web server offers several strategies to prepare the template molecule for the Autodock docking stage (Morris et al., 2009) (Fig.S4.2). The user can input the 3D structure, in all the most common formats (.pdb, .sdf and .mol2 file), without specifying any parameters and the .pdbqt file is generated by using Autodock (Morris et al., 2009) with default parameters, ready to dock.
Advanced users, to gain control on the parametrization of molecule before proceeding to the docking, are funneled to a simplified version of the ACPYPE software program (da Silva and Vranken, 2012), which offers several methods to assign charges to small organic molecule. Indeed, the user can upload a file together with ACPYPE parameters (da Silva and Vranken, 2012) and thus the generated parametrized file (i.e. mol2 file, containing the charges calculated with ACPYPE (da Silva and Vranken, 2012) following the user indications) is used as input for AutoDock (Morris et al., 2009), to obtain the corresponding parameters (.pdbqt). If the molecule is provided in the .sdf format, it is converted to .pdb by OpenBabel (O'Boyle et al, 2011) before run ACPYPE (da Silva and Vranken, 2012).
Additionally, the web server gives also the opportunity to expert users to upload a .mol2, which partial charges can be pre-calculated by an other program/server such as R.E.D.-III (Dupradeau et al., 2010) (http://upjv.q4md-forcefieldtools.org/REDServer/) and Maestro software package (https://www.schrodinger.com/maestro).

The file, containing the 3D coordinates and partial charges, is directly used as input for Autodock (Morris et al., 2009) to obtain the parameters (.pdbqt).
Finally, the users can also upload a pre-generated .pdbqt file, ready to dock. In addition, the user can specify for the input molecules to merge the non-polar hydrogen (as the default option in ADT (Morris et al., 2009)) or to keep all the hydrogens in the model.

Monomer preparation.

- **i) HADDOCK-based approach.** Once the protein/peptide structure is checked and ready to dock, HADDOCK-based pipeline (Van Zundert et.al., 2016 and Dominguez et.al., 2003) requires parameter files with partial charges for the monomers, as well as CNS parameter files. Users can generate them by using the web server, which implements ACPYPE (da Silva and Vranken, 2012) or calculate them by an external web server such as PRODRG (SchuÈttelkopf and Van Aalten, 2004) or ATB (Malde et al., 2011) (Fig.S4.3). The web server includes a virtual library (database) containing the most commonly used FMs (already parametrized), in ready-to-dock, 3D formats, which have been already tested in Cenci et al., 2015 paper.
- **ii) Autodock/Vina-based approach.** The parameterization methods and format file supported are the same as the ones implemented for the AutoDock/Vina receptor preparation section, except for the ADT script used ("prepare_ligand4.py") (Morris et al., 2009) (Fig.S4.4). The user can easily upload the monomers in .pdbqt format, the method for which has already been tested and reported in the literature (Morris et al., 2009).

4.3.1.2 Docking step

i) HADDOCK-based approach. We offer a simplified in-house interface to the HADDOCK software (Van Zundert et.al., 2016 and Dominguez et.al., 2003) that allows the users to intervene in the critical HADDOCK parameters for protein/peptide-ligand docking (Kastritis et al., 2012 and Fiamegos et al., 2011). The web server enables to run the virtual screening of the peptide against the chosen monomers. Moreover, the user can perform multiple ligand docking with up to five monomers together (Fig.S4.5). User must specify the protein residues involved in the receptor-ligand interaction by using a HADDOCK restraint file, which can be easily generated from the HADDOCK web site (Van Zundert et.al., 2016 and Dominguez et.al., 2003) (http://www.bonvinlab.org/software/haddock2.2/generate_air_help/).
Subsequently in our web server, the user can select, optionally, which atom or atoms of the monomer that are putatively involved in the interaction with the protein. These atoms usually correspond to the monomer functional group atoms. This novel improvement has been essential in the screening of the monomers (Cenci et al, 2016 and Cenci et al., 2015), when the latter differ only in their function groups.
ii) Autodock-based approach. The user can perform the virtual screening on the parametrized monomers by using the default Autodock parameters (Morris et al., 2009) or optimized MIPs parameters, giving the user the additional opportunity to upload a custom parameter files. Furthermore, the web server allows advanced users to upload custom Autodock input files (.dpf and .gpf file), which could be generated externally by using MGLTools package (http://mgltools.scripps.edu/). Additionally, the user can also set up some important docking parameters. e.g. the maximum number of evaluations (Fig.S4.6).

iii) Vina-based approach. In this section, the user can perform the virtual screening of monomers or evaluate the rebinding ability of the MIPs by using Vina (Trott and Olson, 2010) as described in the literature (Bates et al., 2016). The web server offers a simplified Vina interface, in which the user can specify the search space and all the docking parameters, which are needed for Vina execution. Furthermore, we offer a set of parameters that have been already extensively tested in the several case studies (Bates et al., 2016) that can be chosen by a simple drop-down menu. If the user does not specify any parameters, the web server will calculate automatically the search space parameters of the receptor (or cavity) by using a VMD (Humphrey et al., 1996) script and default Vina (Trott and Olson, 2010) will be used for the docking (Fig.S4.6).

4.3.1.3 Refining step

This step is an additional step in which we offer a prediction of the initial template: monomer ratio. The user can upload the .pdb file of the receptor and monomer, which are small organic molecules and are parametrized by using the ACPYPE (da Silva and Vranken, 2012) software package. Subsequently, a virtual box containing many copies of the selected monomer around the template molecule is created following the user's specification (by using GROMACS tools) (Pronk et al., 2013). Finally, the monomers are gently relaxed around the template. At the end, the number of hydrogen bond interactions between monomers-template is extracted and the number of monomers interacting with the receptor is shown (Fig.S4.7).

4.3.2 Case of study and Applications

4.3.2.1 HADDOCK based docking step

4.3.2.1.1 Troponin system

4.3.2.1.1.1 Biological Introduction

The fingerprint imprinting (FIP) is a procedure that includes four steps: 1) in-silico digestion of a protein of interest in peptides (7-12 mer); 2) in-silico alignment of the peptide sequences generated with all the protein sequences stored in UniProtKB (blast) ; 3) selection of the most unique peptide (epitope), short and exposed, as template; 4) synthesis of the fingerprint imprinted polymers. Therefore, it is very important to choose one or more key epitopes to use as template to build the MIPs. A good MIP recognizes the peptide that consequently allows the target protein recognition. For this reason, the modeling of the epitope sequences with MIP strategy can be useful and important. Studies have shown that the troponin-T (TnT) and troponin-I (TnI) are correlated with cardiovascular disease. Therefore, the troponin-T (TnT) and the troponin-I (TnI) are used as markers of cardiac dysfunction or damage. In fact, detectable troponin-T was found in 66.5% and clinically significant levels (> 13 ng/l) in 7.4% in the Atherosclerosis Risk in Communities (ARIC) study. In addition, in the Women's Health study (WHS) detectable

troponin levels were found in 30% of high-risk patients especially those with diabetes where TnT was found in 45% (Wierzbicki, 2012).

4.3.2.1.1.2 Aims of this section

- Modelling of MIPs for three peptide sequences i.e. NITEIADLTQK (NK11), NIDALGMEGR (NR11) and AYATEPHAK (AK9) from the protein troponin as templates. In particular, the screening of these three templates against the most commonly used FMs by using a patent protocol based on LeapFrog algorithm and a new HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) based protocol in order to select the high binding monomers to be used in MIP synthesis.

4.3.2.1.1.3 Materials and Methods

The structure of NK11 peptide was extracted from the crystal structure of human cardiac troponin (PDB accession code 1J1E.C) and then relaxed using "minimize structure" tool from the Chimera program (Pettersen et.al., 2004). NR11 and AK9 structures were modeled using a two steps protocol: 1) "build structure" tool from the Chimera program using default parameters for model building and, 2) a few steps of relaxation using "minimize structure" tool.

HADDOCK. The parameters of the studied monomers: 2VP (2-vinylpyridine neutral), 2VP+ (2-vinylpyridine positively charged), 4VP (4-vinylpyridine neutral), 4VP+ (4-vinylpyridine positively charged), AAm (acrylamide), AA (acrylic acid in neutral form), AA- (acrylic acid- negatively charged), acrylonitrile, AMPSA- (2-Acrylamido-2-methylpropane sulfonic acid, negatively charged), TFMAA (2-(Trifluoromethyl)acrylic acid in neutral form), TFMAA- (2-(Trifluoromethyl) acrylic acid, negatively charged), DEAEM (N,N-diethylaminoethyl methacrylate in neutral form), DEAEM+ (N,N-diethylaminoethyl methacrylate, positively charged), EGMP (Ethylene glycol methacrylate phosphate in neutral form), EGMP- (Ethylene glycol methacrylate phosphate, negatively charged, HEM (2-Hydroxyethyl methacrylate), IA (itaconic in neutral form), IA- (itaconic acid negatively charged), MAA (methacrylic acid in neutral form), MAA- (methacrylic acid, negatively charged), VI (vinylimidazole) and VI+ (vinylimidazole, positively charged) were derived using the PRODRG server (SchuÈttelkopf and Van Aalten, 2004). The latter returns topologies and parameter files in various formats, including CNS. These files are needed in the run.cns file. For each of the three templates twenty-two docking experiments were performed (one for each monomer).Every experiment was a molecular docking peptide/monomer. To run docking a run.cns file, containing all the parameters, has to be edited so a number of project-specific parameters such as the number of structures to generate at the various stages, which restraints to use for docking and various parameters governing the docking and scoring can be defined, though many of these parameters have default values. A fundamental change in the run.cns file was the definition of the number of structures in the first two steps (rigid body minimization and semi-flexible SA) and the number of structures to analyse in the third step (explicit solvent refinement). For every experiment, the number of structures for rigid body minimization was 100; the number of structures for semi-flexible SA was 20 and the number of structures for explicit solvent refinement

was 20. Moreover, the AIR file is very important for this problem. The decision was made, in fact, to use all the amino acids of the peptides as active residues for three templates and no residue passive. For the monomers though, it was decided to use the entire monomer, as a residue active, specifying the atom or atoms of the monomer that must interact with the template.

LeapFrog. The virtual library used contains the most commonly used FMs. The virtual library was already created and provided by Leicester University and was used as a database for computational design as previously described in the literature (Chianella et.al., 2002). The LeapFrog algorithm (Tripos Inc., Tripos, St. Louis, MO, USA) was used to screen the library of FMs for their possible interactions with the templates. Monomers giving the highest binding scores represent the best candidates for polymer preparation. There are two different stages to calculate the energies: monomer-template complexation and a system of scoring the complementarities between the monomer and template. In the latter LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA) defines the template as the receptor binding site (using additional site-point matching scores, a system of scoring the receptor and ligand interactions). This two-stage scoring method is used to search a kind of regionality, therefore the "best fit" and the highest binding energy (in kcal/mol) will happen the best score is associated to complementary sites, which are nearby. The LeapFrog parameters (Tripos Inc., Tripos, St. Louis, MO, USA) are standard and they are retrieved from the practical manual (see Fig.3.4). The screening with LeapFrog algorithm (Tripos Inc., Tripos, St. Louis, MO, USA) was performed on each of the three peptide templates.

4.3.2.1.1.4 Results

Over the past decade, the computer power of PCs available in any laboratory has increased, opening new possibilities for simulating aspects of the complex molecular imprinting process. An example of this is the computational method for rational design of MIPs developed in Leicester. This method produces good results but is based on the use of expensive software such as Sybyl program (Tripos Inc.) and LeapFrog algorithm (Tripos Inc., Tripos, St. Louis, MO, USA). For this reason in this chapter, an alternative method to the protocol of Leicester is described. This alternative method is based on use of a state of the art software called HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003), that it is normally used for protein-protein and ligand-protein docking. Here, we will use it to the rational design and analysis of MIPs systems. The methodology was finally implemented in the MIRATE web server.

Modeling of three different peptides. One of the main steps of the protocol developed at Leicester University for the computational design of imprinted polymers is the screening of FMs against the template molecule using the LeapFrog algorithm. LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA) is a component of the Sybyl software package (Tripos) and the algorithm is not free or open source. Therefore, in this section an alternative method to LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA), was designed to the selection of the best FMs for polymer preparation.
Tests. The alternative solution for the screening of the virtual library is to use a HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) based method. HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) software is a method

based on docking to study protein complexes, but this software can be used to the modeling of protein-DNA, protein-RNA, protein-oligosaccharides, protein-ligand and peptide-ligand complexes. Although this tool, for the peptide/monomer docking, produces good results, it may be much slow than the Leicester protocol. Several tests were carried out to find a way of balancing the quality of the results with the computational time required for running HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003). For every experiment, the number of structures for rigid body minimization is 100; the number of structures for semi-flexible SA is 20 instead and the final number of structures for explicit solvent refinement is 20. We have also tried to repeat the same experiments by using default parameters i.e. 1000 as number of structures for rigid body minimization; 200 as number of structures for semi-flexible SA and 200 as the number of structures for explicit solvent refinement. However, we did not see any significant differences in the ranking of the monomers between the two protocols[3], thus we decided to use the fastest protocol (less computational time to generate less structures than the default HADDOCK protocol).

4.3.2.1.1.4.1 Experiments on NK11

Twenty-two docking experiments were performed (one for each monomer). Every experiment performed was a molecular docking peptide/monomer. For each experiment of docking peptide/monomer, only the structure with the best HADDOCK score was analysed. The following table (Table 4.1) shows the results regarding the best structure, on the basis of HADDOCK score, for each type of monomer. In this table (Table 4.1), the monomers are ranked according to their HADDOCK score, and the number of hydrogen bonds with the NK11 template are shown.

Table 4.1. First five best monomers of the screening with HADDOCK. Monomer/Template (NK11) HADDOCK score with the number of hydrogen bonding.

NAME OF MONOMER	HADDOCK SCORE (Kcal/mol)	NUMBER OF HYDROGEN BONDS
IA-	-67.1013	6
EGMP-	-65.8482	8
AMPSA-	-36.8165	6
4VP+	-22.1399	0
4VP	-18.1888	1

Figures 4.4 to 4.6 show the structures of the three best complexes obtained with this approach that are the charged forms of IA, EGMP and AMPSA.

[3] MIRATE web server provides both protocols.

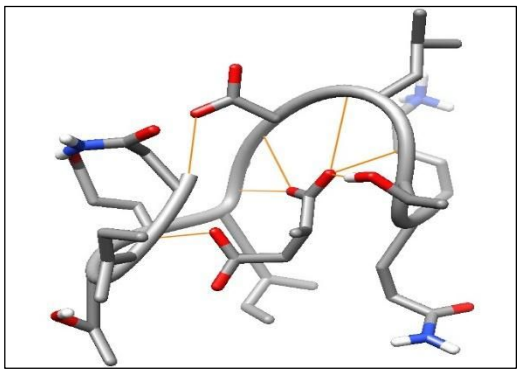


Figure 4.4. Interactions between NK11 and IA- monomer.

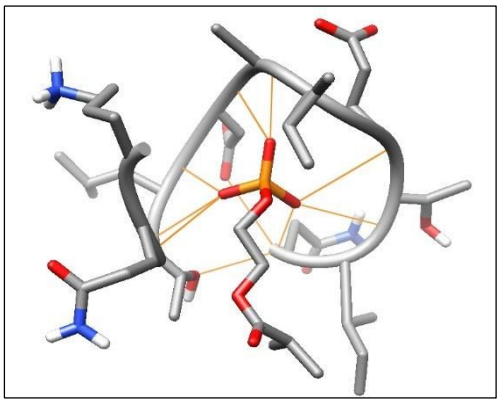


Figure 4.5. Interactions between NK11 and EGMP- monomer.

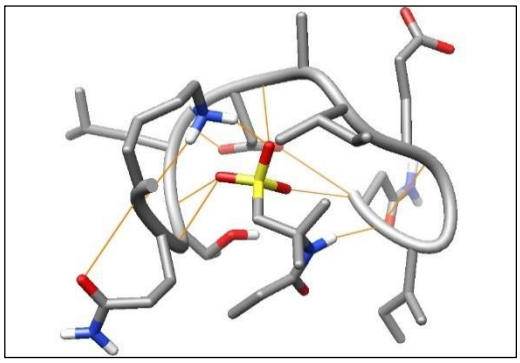


Figure 4.6. Interactions between NK11 and AMPSA- monomer.

These monomers have the best HADDOCK score and form a considerable number of hydrogen bonds. It is worth mentioning that LeapFrog has shown promising results as a predictive tool so any similarity in the modelling results would indicate that HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) has the potential to also be a predictive tool. Below is a summary table (Table 4.2) of the results of HADDOCK (Van

Zundert et.al., 2016 and Dominguez et.al., 2003) and LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA) for the NK11.

Table 4.2. First seven best monomers of the screening with HADDOCK program and LeapFrog algorithm of the virtual library of monomers against NK11.

HADDOCK	LeapFrog
IA-	EGMP-
EGMP-	IA-
AMPSA-	AMPSA-
4VP+	DEAEM+
4VP	MAA-

As it is shown in Table 4.2, the results provided by HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) are very similar to those obtained by using LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA). In particular, for the NK11, HADDOCK provides the same first three monomers as LeapFrog (IA-, EGMP-, AMPSA-) as shown in Table 4.2.

4.3.2.1.1.4.2 Experiments on NR11

Twenty-two docking experiments were performed (one for each monomer). Every experiment was a molecular docking peptide/monomer. For each experiment of docking peptide/monomer, only the structure with the best HADDOCK score was analysed.
The following table (Table 4.3) shows the results regarding the best structure on the basis of HADDOCK score, for each type of monomer. In this table (Table 4.3) the monomers are ranked according to their HADDOCK score, and the number of hydrogen bonds with the NR11 template are shown.

Table 4.3. First five best monomers of the screening with HADDOCK. Monomer/Template (NR11) HADDOCK score with the number of hydrogen bonding.

NAME OF MONOMER	HADDOCK SCORE (Kcal/mol)	NUMBER OF HYDROGEN BONDS
EGMP-	-51.5621	7
IA-	-31.6651	7
AMPSA-	-28.7488	5
VI+	-14.0843	0
EGMP	-13.6044	4

In the following figures, some schemes about the peptide-monomer interactions are reported.

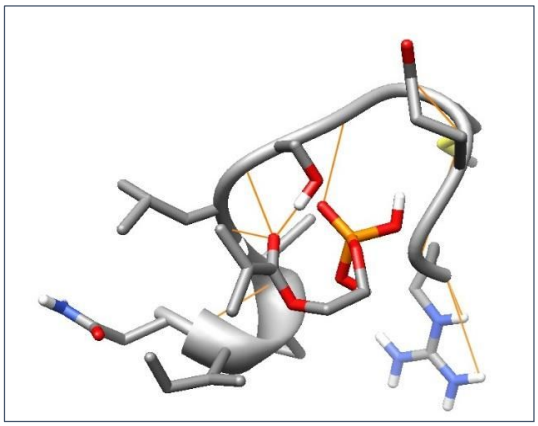


Figure 4.7. Interactions between NR11 and EGMP- monomer.

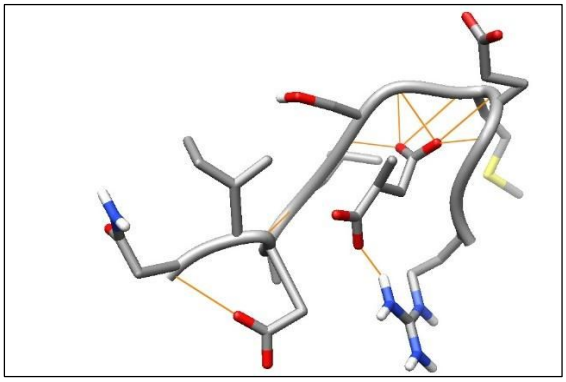


Figure 4.8. Interactions between NR11 and IA- monomer.

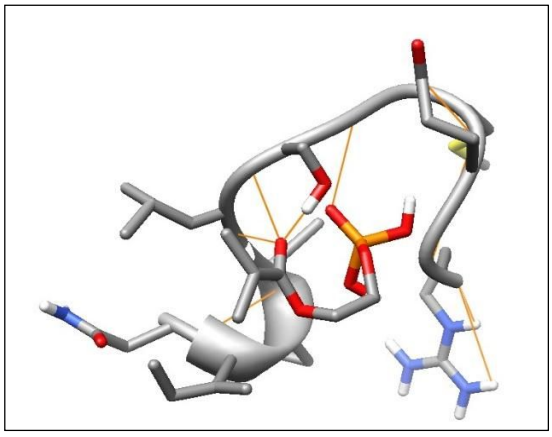


Figure 4.9. Interactions between NR11 and EGMP monomer.

The monomers EGMP-, IA- and EGMP have a good HADDOCK score and form a considerable number of hydrogen bonds.
Table 4.4 summarises the results of HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003)and LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA) for the NR11.

Table 4.4. First five best monomers of the screening with HADDOCK program and LeapFrog algorithm of the virtual library of monomers against NR11.

HADDOCK	LeapFrog
EGMP-	EGMP -
IA-	IA-
AMPSA-	AMPSA-
VI+	DEAEM+
EGMP	MAA

As it is shown in Table 4.4 the results provided by HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) are very similar to those obtained by using LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA). In particular, for the NR11, HADDOCK provides the same first three monomers as LeapFrog (IA-, EGMP-, AMPSA-), in the same order as shown in Table 4.4.

4.3.2.1.1.4.3 Experiments on AK9

Twenty-two docking experiments were performed (one for each monomer). Every experiment was a molecular docking peptide/monomer. For each experiment of docking peptide/monomer, only the structure with the best HADDOCK score was analysed. In Table 4.5 the monomers are ranked according to their HADDOCK score, and the numbers of hydrogen bonds with the AK9 template are shown.

Table 4.5. First five best monomers of the screening with HADDOCK. Monomer/Template (AK9) HADDOCK score with the number of hydrogen bonding.

NAME OF MONOMER	HADDOCK SCORE (Kcal/mol)	NUMBER OF HYDROGEN BONDS
IA-	-71.0555	9
EGMP-	-61.8783	8
AMPSA-	-32.0393	6
MAA-	-30.8658	5
AA-	-29.5291	5

In the following figures, some schemes about the peptide-monomer interactions are reported.

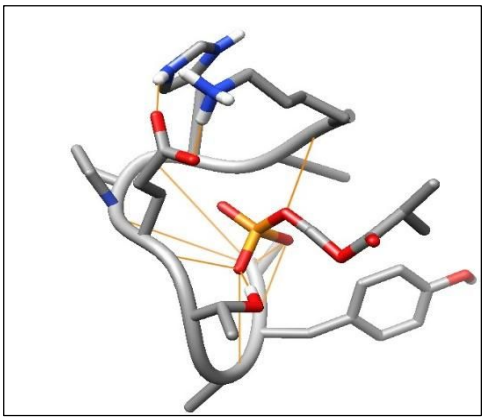


Figure 4.10. Interactions between AK9 and EGMP- monomer.

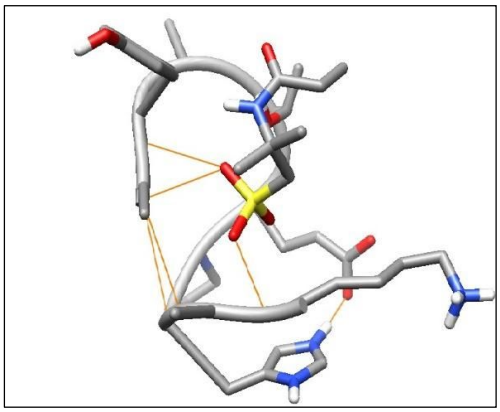


Figure 4.11. Interactions between AK9 and AMPSA- monomer.

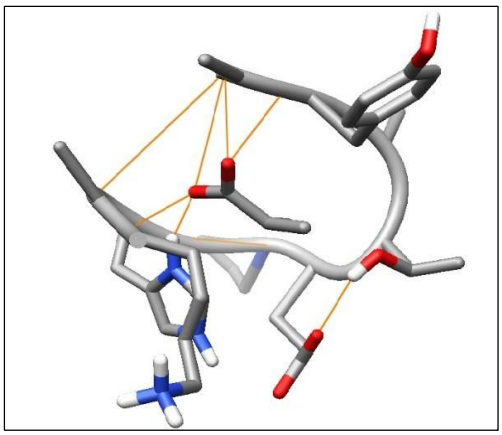


Figure 4.12. Interactions between AK9 and MAA- monomer.

The monomers EGMP-, AMPSA-, MAA- have a good HADDOCK score and form a considerable number of hydrogen bonds.
The results of screening of HADDOCK for AK9 compared these with the results obtained with LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA). Table 4.6 summarises the results of HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) and LeapFrog for the AK9.

Table 4.6. First five best monomers of the screening with HADDOCK program and LeapFrog algorithm of the virtual library of monomers against AK9.

HADDOCK	LeapFrog
IA-	EGMP-
EGMP-	AMPSA-
AMPSA-	IA-
MAA-	TFMAA-
AA-	MAA-

As it is shown in table 4.6, the results provided by HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) are very similar to those obtained by using LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA). In particular, for the AK9, HADDOCK provides the same first three monomers as LeapFrog (IA-, EGMP-, AMPSA-) as shown in Table 4.6. Also, MAA- monomer is found to be in the first five positions of the comparison in Table 4.6.

4.3.2.1.1.5 Discussion and Conclusions

Here we have tested our HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) based methods and LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA) for the selection of monomers. In general, we can appreciate that, although the algorithms underlying both protocols are different, and that the energy values cannot be compared as an absolute value, the results obtained are very similar indicating that HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) based method represent a valid alternative for the designing of MIPs and for performing virtual screening of monomers. With our protocol, we were able to select high affinity monomers for all templates as potential monomer candidates for imprinting. In particular, our predictions indicate itaconic acids as the best suitable monomer for MIPs synthesis. Subsequently our collaborators synthesized high affinity synthetic receptors for detection of cardiac Troponin I peptides. The MIPs affinity for their templates was evaluated by isothermal titration calorimetry (ITC). During the titration, the binding event is measured as heat changes (μJ). Both a favorable enthalpic contribution ($\Delta H^0<0$) and an overall negative free energy change ($\Delta G^0<0$) was detected for the three MIP/peptide titrations, showing spontaneous peptide binding in the experimental conditions. Nanomolar affinities were

calculated, suggesting the MIP materials showed a performance comparable to antibodies and properties similar to other micron and nano-gel sized MIPs. The results and all the experimental data have been extensively discussed and published (Cenci et.al., 2016). All the docking experiments can be easily repeated by using our MIRATE web server (http://mirate.di.univr.it/dock/). This project was successfully performed by a profitable collaboration with Prof. Alessandra Maria Bossi.

4.3.2.1.2 Hepcidin system

4.3.2.1.2.1 Biological Introduction

Diseases related to an altered iron homeostasis which has a high prevalence in the population have led to the development of biosensors for efficient measurement of new or established molecular markers for the control of this chronic and degenerative pathologies. About 2.5-10% of the European population have iron overload diseases, such as hereditary hemochromatosis (HH). HH is a disease that causes unregulated absorption of iron from the gut, resulting in an iron overload. The most common form of HH is caused by mutations in the HFE gene. More than 500 million individuals in the world suffer of iron deficiency anemia (IDA). Hepatocytes in the bloodstream release the peptide hormone hepcidin in order to regulate the efflux of iron in the plasma. For this reason the hepcidin is very important to regulate systemic iron homeostasis in mammals, as well as it is linked with basic pathways like inflammation. Hepcidin is a circulating peptide hormone. In the liver, by means of proteolytic cleavage, the mature hepcidin is built. In addition there are two different types of hepcidin, with a not clear biological role, hepcidin-20 and hepcidin-22; they main characteristic is that they don't have the 5 and 3 residues at the N-terminus respectively. The only "bioactive" form has 25 residues and its structure has a hairpin folding, constrained by four disulphide bridges. Hepcidin-25 is the "master regulator" of the iron metabolism by binding to the iron channel ferroportin (Fpn). This binding results in the internalization of the hepcidin-ferroportin complex with subsequent degradation of ferroportin, producing a block of the metal ion efflux in the blood. Ferroportin is a transmembrane protein that transports iron from the inside of a cell to the outside of it and it is localized in any cell type that is involved in iron pathways. Inhibiting ferroportin shuts off the iron transport out of the cells such as macrophages, duodenal mucosa and hepatocytes. In effect thanks to the recycle of old red blood cells, the macrophages are the main source of iron while the duodenal mucosa accounts for about 10% of the iron daily required by the bone marrow. In parenchymal tissues such as the liver, the iron is storage. Therefore, the analysis of hepcidin-25 levels in the human body is a powerful tool for diagnosis and prognosis in many human diseases. In addition, blood tests can be done to evaluate levels of the hepcidin to study the problem of doping among athletes since there is a correlation between iron metabolism and erythropoietic activity. To conclude, one important problem is to find a method that recognizes and distinguish the 5 residues at the N terminal typical of the "bioactive" hepcidin-25 from the other two isoforms, hepcidin-20 and hepcidin-22, lacking of 5 and 3 residues at the N-terminus respectively. Then the 5 residues at the N terminal are important for determining the quantity of mature hepcidin present in the circulation (Anderson and McLauren, 2012).

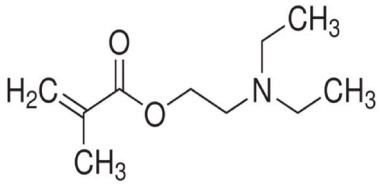
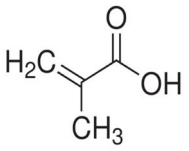
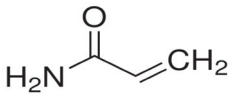
4.3.2.1.2.2 Aim of this section

- Molecular modelling of MIPs for hepcidin-25 human peptide as a template by using HADDOCK based method for monomer selection.

4.3.2.1.2.3 Materials and Methods

The template used is hepcidin-25 peptide downloaded from the PDB website with code 2KEF. The monomers used are Diethylaminoethyl methacrylate (DEAEM), Methacrylic acid (MAA) and Acrylamide (AAm). The structures of the monomers are shown in Table 4.7.

Table 4.7. Structures of monomers used for hepcidin-25 peptide as the template

Diethylaminoethyl methacrylate (DEAEM), positively charged monomer	
Methacrylic acid (MAA), negatively charged monomer	
Acrylamide (AAm), neutral monomer	

The monomers were built by using PRODRG (SchuÈttelkopf and Van Aalten, 2004) server. This program returns topologies and parameter files in various formats, including CNS. These files are needed in the run.cns file. To run docking a run.cns file, containing all the parameters, has to be edited. Therefore, a number of project-specific parameters such as the number of structures to generate at the various stages, which restraints to use for docking and various parameters governing the docking and scoring can be defined, though many of these parameters have default value. Instead, the definition of the AIR is very important for this problem. It was decided, in fact, to use the 5 amino acids at the N terminal as active residues for the hepcidin template. For the monomers though, it was decided to use the entire monomer, as a residue active, specifying the atom or atoms of the monomer that must interact with the template.

Then the two docking tests were performed:

1. Docking of the human hepcidin-25 (target) with 2 molecules of DEAEM (positively charged) and two molecules of AAm (neutral)

2. Docking of the human hepcidin-25 (target) with 2 molecules of MAA (negatively charged) and two molecules of AAm (neutral)

4.3.2.1.2.4 Results

In this section, the modelling of MIPs is presented using our HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) procedure. The formation of the complex between hepcidin-25 peptide, used as template, and a number of FMs (DEAEM, MAA and AAm) is proposed. The goal was to find a right combination of monomers to suit the N-terminus of the hepcdin-25, providing information on how to build an appropriate MIPs able to recognize the N-terminus and therefore the hepcidin peptide.

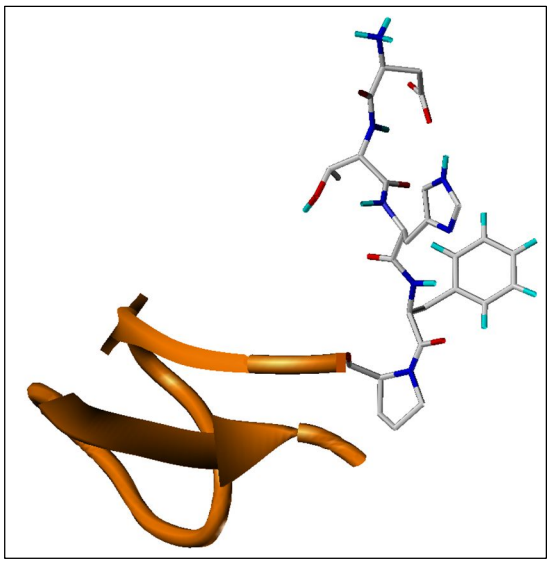


Figure 4.13. Cartoon representation of hepcidin with stick representation of its N-terminal (DTHFP)

MIPs construction, in the pre-polymerisation mixture, requires that only the region of the monomer containing a functional group, should interact with the template, the crosslinker holds the complex in place during polymerisation. Thus, a tool able to allow the introduction of constraints on the orientation of monomers during the docking procedure is needed. The HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) program, based on knowledge–based scoring functions, is indeed the key to face this challenge. Several approaches have been tried using HADDOCK software to solve the modeling of the hepcidin peptide. At first, simple docking tests have been performed using one monomer (DEAEM, MAA, AAm) and hepcidin-25 as template. Docking experiments were designed using a different number and type of monomers together with the target molecule, in order to mimic the complex formation between the template and FM present in the pre-polymerisation mixture. Two main tests were designed, able to provide exact information about the construction of a MIPs for recognizing the N terminal in the hepcidin peptide.

1. Docking of the human hepcidin-25 (target) with two molecules of DEAEM (positively charged) and two molecules of AAm (neutral)

2. Docking of the human hepcidin-25 (target) with two molecules of MAA (negatively charged) and two molecules of AAm (neutral)

The aim of our study is to find a right combination of monomers to model the N-terminus of the hepcidin-25, so the crucial point of these experiments is the definition of the spatial restraints (AIR file) that specifies which amino acids and atoms can interact with each other during the docking process. For this task, the five amino acids at the N terminal are considered as active residues for the hepcidin template. Each monomer was then considered as an active residue. This strategy was used for all the experiments described in this section.

4.3.2.1.2.4.1 Hepcidin: first test

The first test is based on the docking of the hepcidin with two molecules of DEAEM and two molecules of AAm. At the end of the process of docking, HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) gives an energetic score to every structure.
The scoring, i.e. HADDOCK score, is obtained according to the weighted sum of the following terms:

- E_{vdw}: van der Waals energy
- E_{elec}: electrostatic energy
- E_{dist}: distance restraints energy
- E_{sani}: direct RDC restraint energy
- E_{vean}: intervector projection angle restraints energy
- E_{dani}: diffusion anisotropy energy
- E_{cdih}: dihedral angle restraints energy
- E_{sym}: symmetry restraints energy
- BSA: buried surface area
- dE_{int}: binding energy
- E_{desol}: desolvation energy

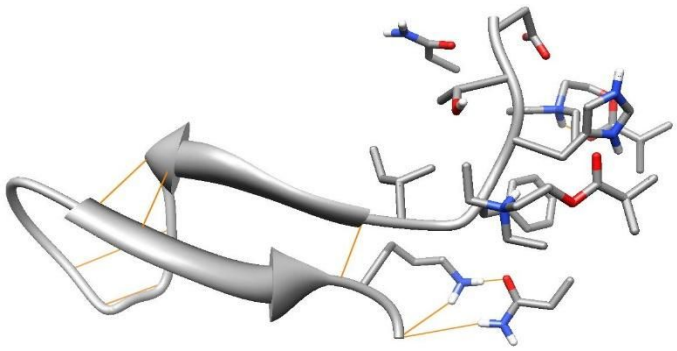
The structure with the lowest value of HADDOCK score may indicate the complex with the lowest energy. The peptide-monomers complexes produced by HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) based method were grouped into clusters. The cluster are created by using a 7Å RMSD cut-off and requiring a minimum of four structures per cluster. This cut-off value is a good compromise between the number of clusters products and the number of structures for each cluster. The analysis of the number of structures in a cluster, allows us to understand if that type of the complex found is reliable and feasible. For these reasons, the three most populated clusters were examined and the best structure for each of these three clusters was analysed. Therefore,

for each test, three structures were examined. This strategy was used for all the experiments described in this section.
Table 4.8 shows the results of the first test. In this table, the names of the three structures with the corresponding energy values and hydrogen bonds are shown.

Table 4.8. Hepcidin: first test. Monomers/Template HADDOCK score with hydrogen bonding distance. In gray, green and pink are the information about cluster1, 2 and 3 respectively.

NAME COMPLEX	NUMBER CLUSTER	HADDOCK SCORE (a.u.)	NAME MONOMERS	ATOMS MONOMER	ATOMS TEMPLATE	LENGTH (Å) H_BOND INTERACTION
126W.pdb	1	-59.2420	DEAEM			
			DEAEM			
			AAm	HAB OAA	THR25 OXT LYS24 HZ1	2.192 1.717
			AAm			
46W.pdb	2	-54.8660	DEAEM			
			DEAEM	OAB	HIS3 HD1	1.662
			AAm	HAB	THR2 O	2.225
			AAm			
156W.pdb	3	-47.1522	DEAEM	HAC	ASP1 OD1	1.866
			DEAEM			
			AAm	OAA	HIS3 HD1	1.679
			AAm	HAB	THR2 O	2.210

Figure 4.14 shows the representative structures for each of the most populated clusters. In Figure 4.14 the atoms in red are oxygen; in blue are nitrogen; in white are hydrogen and grey is shown the carbon. The hydrogen bonds are shown in orange.



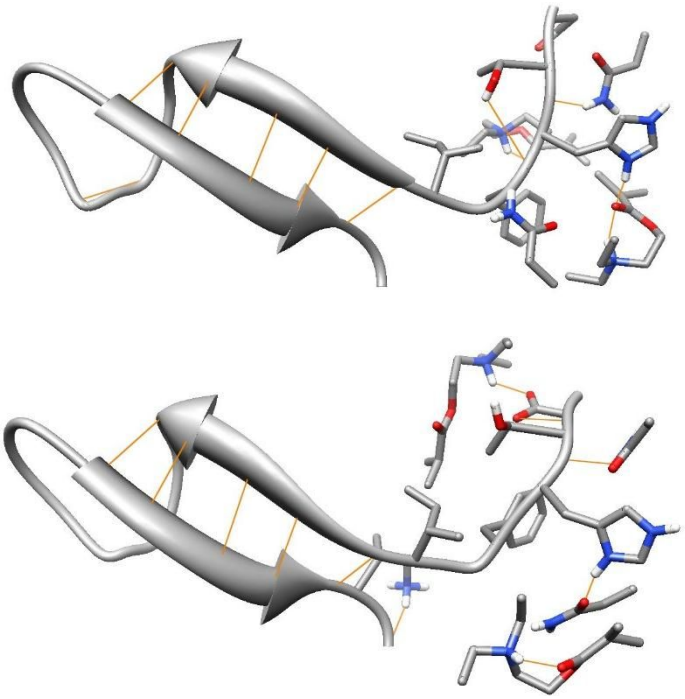


Figure 4.14. Hepcidin: first test. Representative structures for each of the most populated clusters. top) first cluster; middle) second cluster; bottom) third cluster.

4.3.2.1.2.4.2 Hepcidin: second test

The second test is based on the docking of the hepcidin with two molecules of MAA and two molecules of AAm.
Table 4.9 shows the results of the second test. In this table, the names of the three structures with the corresponding values for energy and hydrogen bonds are shown.

Table 4.9. Hepcdin: second test. Monomers-Template HADDOCK score with hydrogen bonding distance. In gray, green and pink are the information about cluster1, 2 and 3 respectively.

NAME COMPLEX	NUMBER CLUSTER	HADDOCK SCORE (a.u.)	NAME MONOMERS	ATOMS MONOMER	ATOMS TEMPLATE	LENGTH (Å) H_BOND INTERACTION
137W.pdb	1	-67.9775	MAA	OAB OAA	ASP1 HT3 ASP1 HT3	1.703 2.347
			MAA	OAA OAA OAB OAB	HIS3 HD1 HIS3 HN THR2 HG1 HIS3 HN	1.675 1.759 1.718 2.623
			AAm	OAA	PHE4 HN	2.106
			AAm	HAG	THR2 O	2.229
182W.pdb	2	-72.0490	MAA	OAB	LYS24 HZ2	1.723
			MAA	OAB	HIS3 HD1	1.973
			AAm	OAA	THR2 HN	1.709
			AAm			

150W.pdb	3	-67.8225	MAA	OAA OAB	ASP1 HT1 HIS3 HE2	1.698 1.710
			MAA	OAA OAB	LYS24 HZ1 LYS24 HZ1	1.786 1.863
			AAm	OAA	PHE4 HN	2.094
			AAm	OAA OAA	THR2 HN ASP1 HT3	1.764 1.709

Figure 4.15 shows the representative structures for each of the most populated clusters.

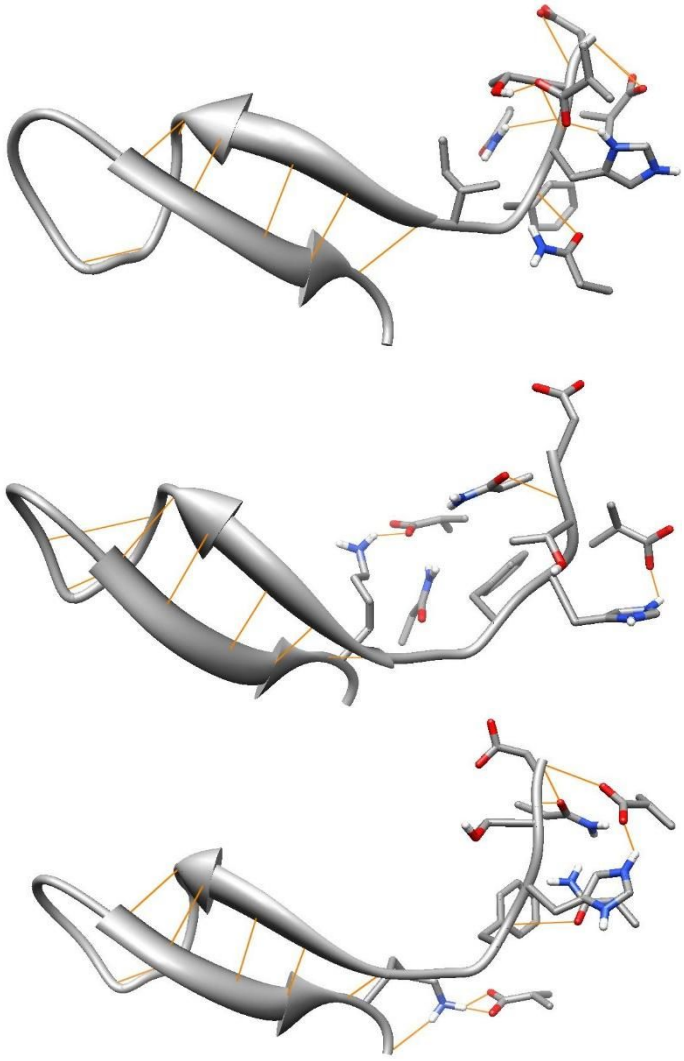


Figure 4.15. Hepcdin: second test. Representative structures for each of the most populated clusters. top) first cluster; middle) second cluster; bottom) third cluster.

Analysing the monomers/template complexes obtained in the two tests we can conclude that the monomers achieve an optimized conformation with the template with a low energy of the overall system. In particular, the second test provides structures with lower values of HADDOCK score compared to the structures produced by the first test. Indeed, the structures resulting from the second tests form more hydrogen bonds than the resulting

structures of the first test. This might be due to possible double interaction of the molecule MAA (both with acid aspartic both and histidine).

4.3.2.1.2.5 Discussion and Conclusions

In this chapter we use a new HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) based protocol for hepcidin-25 human peptide as a template for monomer selection as a new tool in MIP design. HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) was investigated for MIP design as a novel and alternative tool to Sybyl (Tripos Inc.). This study stems from the desire to find a free alternative computational method to the Leicester protocol that relies on purchasing Sybyl (Tripos Inc.) (approximately £4,000 per annum). HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) software allowed the introduction of constraints arising from the knowledge of the chemical-physical properties during the docking procedure setting some parameters. The tests were performed with the default parameters and here, the N terminus of hepcidin-25 was docked with a combination of charged (either DEAEM or MAA) and neutral monomers (Aam): to mimic the pre-polymerization solution two monomers of a same charge and two neutral monomers were funneled through the HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) software. At the end of the simulations the HADDOCK score values, that indicate the strength of the complex, showed a stronger interaction for MAA/template (−68.0 a.u.) respect to DEAEM/template (−59.2 a.u.). Our predictions indicated MAA as best monomer for the MIPs synthesis. Finally, our collaborators manufactured MIPs with high selectivity and affinity for N terminus of the hormone. The project was carried out in close cooperation with Prof. Alessandra Maria Bossi (Cenci et.al., 2015). All the docking experiments can be easily repeated by using the MIRATE web server (http://mirate.di.univr.it/dock/).

4.3.2.2 Autodock/Vina based docking step

In different and independent systems, the "Autodock" (Morris et al., 2009) based approach was adopted to discover the monomer-template interactions in term of binding score and detection of hydrogen bond formation, location and directionality. The FMs, showing highest binding score and capable of forming strong complexes with theophylline and caffeine (templates) were in agreement with the experimental data at the pre-polymerization and rebinding phase (Manuscript in preparation, Bates et al., 2017).[4] Beyond the encouraging results in the ranking of the monomers against a whatsoever type of template, the web server allows also the screening of "virtual imprinted receptor" for rebinding of target analyte as well secondary structures. This novel technique based on Vina algorithm (Trott and Olson, 2010) and already extensively described in literature with all its advantages (Bates et al., 2016), enables to assess the rebinding ability of the MIPs, in consideration of the cumulative steric effect of multiple FMs. The latter has been successfully applied in three cases, either confirming the experimental data available and aiding the synthesis of more selective MIPs. The case of applications of this method were:

[4] We cannot provide more details for privacy reason, however, MIRATE web server offers all the components needed to perform these kind of calculations, hiding the laborious tasks which would otherwise need to be manually done by user.

1) the detection of tetrahydrocannibinol using a Catechin–hydrate MIP with acrylamide as monomers; 2) the detection of Bisphenol A and Bisphenol F using a Bisphenol A MIP with 4VP as the monomer and 3) the detection of hypoxanthine by a selection of theophylline MIPs. The cases studies are extensively described in Bates et al., 2016.

4.3.2.3 GROMACS based refinement step

In this section, the monomers giving the highest binding score and capable of forming the strongest complexes with the template are usually selected as candidates for the polymer preparation. Subsequently, the template molecule and multiple copies of the monomers chosen are assembled in a virtual box and the energy of the system is minimized. The parameters applied to the refining step, are described in section 4.3.2.3.1.3, but these latter may be changed in order to mimic the actual conditions encountered in the real polymerisation or rebinding experiments in lab. At the end of this minimization process, an analysis of the H bonds interaction between the monomers and the template is done. The purpose of the refining step is to extract information about the idea stoichiometric ratios of monomers to template. After this analysis, only the monomers that interact with the template are extracted together to the template. The type and quantity of the monomers participating in the complex determine the ratio of the template and monomers used for the polymer preparation. In this section we present an example of calculation of the melamine (template): itaconic acid (monomer) stoichiometric ratio. The content of this chapter has been published in Bates et al., 2017.

4.3.2.3.1 Melamine

4.3.2.3.1.1 Biological Introduction

Melamine, a highly nitrogenous compound, is a heterocyclic triazine. It is used extensively in the synthesis of melamine formaldehyde resin which is, in turn, used in the fabrication of everyday products such as laminates, fabric coatings, commercial filters, glues and adhesives in addition to its general use in food containers and various common kitchenwares (Rovina and Siddiquee, 2015). Given its close proximity to food stuffs, the need to accurately detect and monitor it is paramount, especially when melamine combines with cyanuric acid (National Toxicology Program, 1983 and Panuwet et al., 2010). The hydrogen bonded complex of the two molecules is highly insoluble and precipitates in the kidneys to form stones which can lead to potentially fatal kidney failure (Skinner et.al., 2010). Due to the aforementioned nitrogen content, counting for two thirds of its molecular mass, adulteration of dairy products with melamine has been employed as a method of artificially augmenting the apparent crude protein content (Lynch and Barbano, 1999) and as a cheap food supplement in cattle feed (Venkatasami and Sowa, 2010). Contamination of infant formula and pet food with melamine has caused a global scandal following the deaths of a number of babies and the hospitalisation of several thousand more (Gossner et.al., 2009). In 2004 the World Health Organisation (WHO) established a tolerable daily intake (TDI) of melamine of 0.2 mg/kg of body mass, a reduction from the previous level of 0.63 mg/kg set by the American FDA based on data collected in 1983 (Rovina and Siddiquee, 2015). As a result the call was issued for more discriminating detection methods for melamine. For this reason, great efforts have been

made in recent years to develop reliable methods for the detection for this molecule. Exhaustive research has been performed on the application of aptamers, immunoassays, colorimetric and electrochemical sensors for analysis of melamine (Rovina and Siddiquee, 2015 and Li et.al., 2015). Unfortunately, these approaches suffer from high price and limited dynamic and thermal stability of specific binders used in the assays and sensors. It is logical then that the application of MIPs for the detection of melamine has been attempted using several different imprinting and polymerisation methods.

4.3.2.3.1.2 Aim of this section

- to determine the ideal melamine (template): itaconic acid (monomer) stoichiometric ratio by using a novel MM method based on the software suite 'GROMACS' (Pronk et.al., 2013).

4.3.2.3.1.3 Materials and Methods

The structures of melamine and itaconic acid molecules were downloaded from ZINC database (Irwin et.al., 2012) and then minimized using tools from the Chimera program (Pettersen et al., 2004). A virtual library of the most common FMs was screened against the melamine by using LeapFrog algorithm. Itaconic acid exhibited high binding score toward melamine, thus it was used as monomer for the refinement step. The melamine molecule was placed in box (3.6×3.6×3.0 nm) with itaconic acid monomers. The resulting system consisted of ca. 2175 atoms. The AMBER99sb force field was applied (Hornak et.al., 2006). Parameters for itaconic acid and melamine molecules were derived using the ACPYPE program using default parameters (da Silva and Vranken, 2012). The system was energy minimised and then two cycles of geometry optimisation (GO) were applied using GROMACS 4.5.5 (Pronk et.al., 2013). The first GO was carried out at 300 K and 1 atm by performing 1 ns of gradual annealing, thus the temperature was gradually increased from 0 to 300 K. The second GO simulation was realised at 300 K and 1 atm by performing 200 ps. The system was geometry optimised in two cycles comprising 800 steps of steepest descent followed by 3000 steps of conjugate gradient. During the GO phases, Berendsen thermostat and barostat (Berendsen et al., 1984) were applied to control the temperature and pressure, respectively. The LINCS algorithm (Hess et.al., 1997) as used to constrain all bond lengths involving hydrogen atoms and an integration time step of 2 fs was used. Periodic boundary conditions (PBC) were applied. Long-range electrostatic interactions were treated using the particle mesh Ewald (PME) method (York et al., 1994). The cut-off radius for the real part of the electrostatic interactions, as well as for the van der Waals interactions was set to 1 nm. At the end of these GOs, the number of hydrogen bond interactions between melamine and itaconic acid molecules was extracted by using Chimera software (Pettersen et al., 2004).

4.3.2.3.1.4 Results and Discussions

Melamine-itaconic acid stoichiometric ratio of 6 was determined by our approach (see 4.3.2.3.1.3 Materials and Methods).

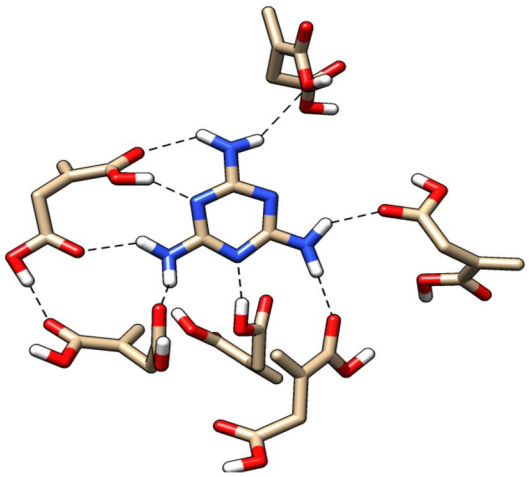


Figure 4.16. Melamine-itaconic acid complex.

To validate our results we repeated the experiment using the established SA process based on the Sybyl (Tripos Inc., St. Louis, MO, USA) molecular modelling suite (Piletsky et al., 2001). The Sybyl protocol involves several steps as follows. A virtual cubic box containing one melamine molecule was packed with itaconic acid monomers until saturation capacity. The box was then reduced to its minimal dimensions to guard against excessive expansion of the mass as a result of repulsion between the molecules during the simulation. This box was then heated to 600ºK and temperature was reduced in four 25 femtosecond steps to 300ºK during which time the initially high mobility of the molecules had decreased to a relaxed state of lowest energy and thus permitting the formation of a complex with the template molecule. Once the simulation was complete, the stoichiometric ratio was determined visually by observing the number of monomers bonded to the template molecule. Even in this case, the melamine-itaconic acid stoichiometric ratio of 6 was determined, with both complexes (calculated by GROMACS (Pronk et.al., 2013) and Sybyl (Tripos Inc., St. Louis, MO, USA)) showing identical bonds for each. This ratio of 6 with respect to the use of a carboxylic dipole is also independently confirmed by a third method (Wang et.al., 2015). Thus, the novel determined parameters for refinement step protocol based on the GROMACS open source package were doubly confirmed. The two rational approaches i.e. Sybyl (Tripos Inc., St. Louis, MO, USA) and GROMACS (Pronk et.al., 2013) are based on approaches of MM. The purpose of these experiments is to optimize the arrangement of FMs around the template combining the results of two slightly different approaches:

a) Sybyl protocol (Tripos Inc., St. Louis, MO, USA) performs a relaxation of the monomers around the template molecule by using SA (Gould et al., 1988) and this method has been widely described in several papers (Subrahmanyam et al., 2012; Subrahmanyam et al., 2001; Piletsky et al., 2001 and Chianella et al., 2002, Monti et al., 2006 and Lv et al., 2008, Karim et al., 2005). Generally speaking, SA is mainly employed for two goals: 1) to surmount energetic barriers of the system, searching conformations with low energy

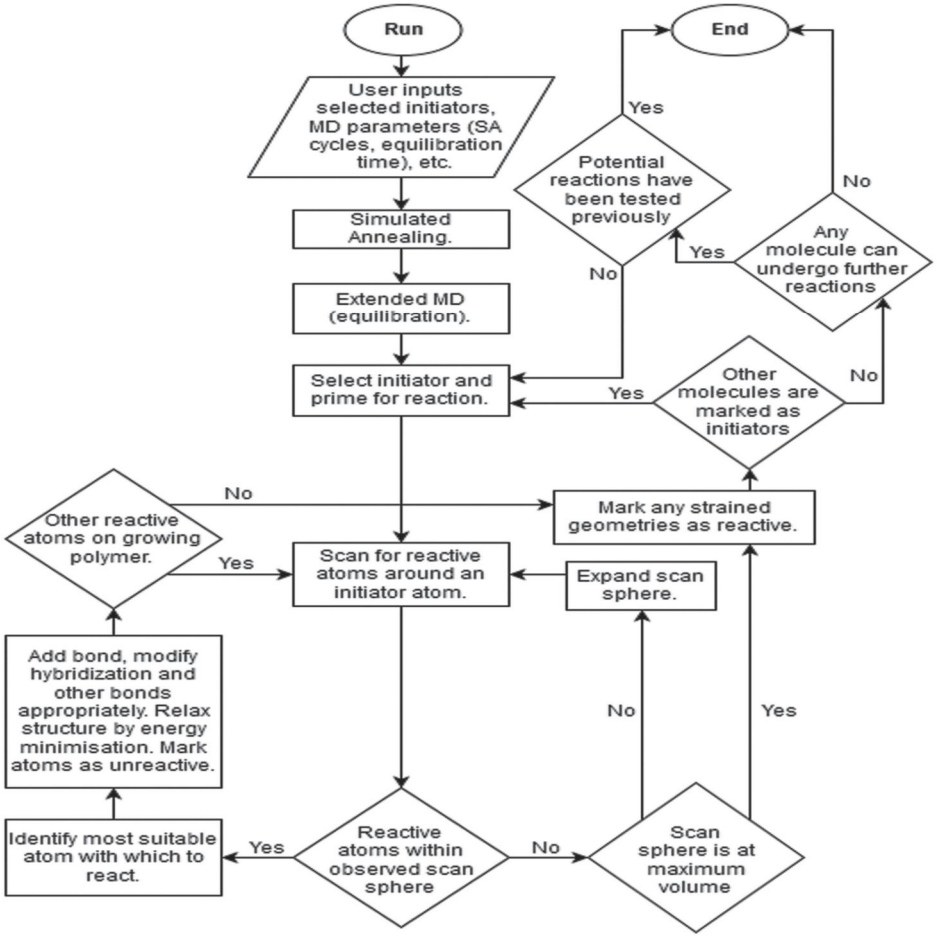
in a fast way; 2) to bring the system temperature distribution to the final target temperature (300 K).
b) A novel computational protocol that involve the use of a state of the art software (GROMACS (Pronk et.al., 2013)), is presented (see section 4.3.2.3.1.3). In detail, several cycles of minimization are performed to reach an initial stable system. Subsequently, a relaxation phase is carried out. This strategy enables the geometric optimization of the monomers around the template as in the Sybyl protocol (Tripos Inc., St. Louis, MO, USA). Indeed, the method is widely employed in many complex systems (e.g. proteins in water solvent and other sensitive molecules) (Karlsson et al, 2009; Chong et al., 1999; Piccoli, et al., 2014 and Musiani et al., 2010) and it allows to overcome energy barriers, bringing the monomers-template complex to a global minima at the target temperature (i.e. 300 K).
However, the use of a computationally determined ratio alone, which is calculated under ideal conditions, has been reported, for our system under investigation, to produce an inferior MIP with a high incidence of non-specific and heterogeneous binding behaviour caused by the disruption of the majority of the template-FM complexes by the solvent and cross-linker (Schauperl and Lewis, 2015). For this reason, an excess of monomer (itaconic acid) as a 'factor of safety' was added to the calculated ideal melamine: itaconic acid ratio and a successfully MIPs for melamine was manufactured (Bates et al., 2017).
To conclude, we developed and applied a computational protocol based on a publicly available software (GROMACS) (Pronk et al., 2013) that is able to output a prediction. In particular, we decided to lower the computational requirements by simplifying the model, estimating how the monomers bind to template and the monomer: template ratio in short time. The combination of these two approaches (GROMACS and Sybyl) could be advantageous as a prediction tool; helping in the design of the MIPs.
Finally, MIRATE web server can be used to calculate the template: monomer stoichiometric ratio.

4.4 Implementation of a new algorithm to *in silico* synthesis of MIPs

The results described in this chapter have been published in Cowen et al., 2016 paper.

4.4.1 Introduction

Approaches to the design of MIPs have fluctuated in popularity through time, but the use computational molecular modeling has remained in use since its initial adoption in the late 1990s (Castro et al., 1998), with the 2001 publication by Piletsky et al. establishing the use of computational screening as an efficient method of functional monomer selection (Piletsky et al., 2001). FMs (polymerizable molecules with functionalities for association to the template) are ranked based on their potential for forming a stable monomer-template complex, and thus forming a strong association in the prepolymerization mixture and a high affinity in the final imprinted polymer. The successes documented using this technique are widespread, but the approach is not without its faults. Problems of representation in simply measuring the monomer-template interaction in the absence of solvent molecules and cross-linker monomer (which typically have the greatest contribution to the mass of the final polymer), and the assumption of similar relative affinities between the prepolymerization mixture and the resultant polymer have raised concerns in the past. However, attempts to fully develop more realistic models are rare, and mostly rely on relatively simple modifications to the existing protocol. We present here a new algorithm (see Scheme 4.1), capable of bringing a random arrangement of molecules to an equilibrium arrangement and simulating polymerization. The algorithm mimics a radical polymerization mechanism and produces, *in silico*, imprinted polymers which can be analyzed by a variety of methods to determine appropriate prepolymerization mixture composition for the optimal production of synthetic receptors. The resulting polymers are believed to be the most realistic atomistic model MIPs thus produced, and the results of analysis demonstrate that this approaches has significant advantages in accuracy over existing techniques, giving strong correlation with empirical data in predictions of affinity and selectivity, in scenarios where previous methods can be shown to be insufficient.



Scheme 4.1. A simplified flowchart of the polymerization algorithm. The user inputs the SA temperature maximum and minimum, number of cycles, and gradient. The user is then asked how long the equilibration period (dynamic simulation under conditions) should be, and to give the reference number of selected initiator molecules, or alternatively to select all reactive molecules to function as potential initiators, as under light-induced polymerization or in approximating one small area of a larger polymerization mixture (scheme retrieved from Cowen et al., 2016 paper).

4.4.2 Materials and Methods

All modeling was carried out by using the Sybyl 7.3 software (Tripos Inc) package. Experiments were performed based on the polymerization mixtures in the methacrylic acid (MAA), hydroxyethyl methacrylate (HEM) and 2-vinylpyridine (2VP) used by Piletsky et al. in their design of a MIP for (–)-ephedrine (Piletsky et al., 2001). Using one example of each molecule, the alkene atoms of the functional monomer were named ‘alpha’ or ‘beta’ depending on the favorability of reaction (assignment of other potentially reactive atoms is performed automatically during polymerization). An additional script was used to build a sphere of randomly arranged molecules in the numbers given in Table 4.10, following the composition of the original article, equal to 5 template molecules. Ethylene glycol dimethacrylate (EGDMA) was used as the cross-linker and chloroform as solvent. System A included MAA as functional monomer, B HEM, and C 2VP.

Table 4.10. The numerical composition of the prepolymerization systems observed.

ID	template	functional monomer	cross-linker	solvent
A	5	50	104	216
B	5	50	149	309
C	5	50	124	256

The volume of each system was given by equation 4.2:

$$\text{Volume (nm}^3\text{)} = \sum_i^n \frac{n_i M_{r_i} 10^{21}}{n_T N_A p_i}$$

Equation 4.2. Volume of each system.

Where the index *i* refers to the component (template, functional monomer, crosslinker or solvent), thus n_i is the number of moles of the component in the original article, its relative mass and p_i its density, while n_T is the number of moles of template and N_A is the Avogadro constant. This volume was used to determine the dimensions of the box within which the system would be equilibrated, each side having a length the cube root of the volume.

The process of system equilibration is controlled from a series of prompts on initiating the algorithm. The first set center on SA and MD; in these experiments each system proceeded through 5 cycles of stepwise annealing from 800 K to 300 K with 100 K intervals and 10,000 femtoseconds at each temperature before 2 nanosecond equilibration periods at 300 K. All MD simulations were carried out using the NVT canonical ensemble (Nauchitel, 1981) with the Tripos force field (Clark et al., 1989), MMFF94 charges (Halgren, 1996) and a non-bonding interaction cut-off distance of 8Å. The polymerization algorithm is written in Sybyl programming language (SPL) and designed for use with the Sybyl program (Tripos Inc). The major loop runs through the molecules assigned as initiators, or all molecules with atoms named as reactive. In the initiator molecule or growing polymer, the reactive atoms scan out in concentric circles for reactive molecules, in these experiments in increments of 10 pm, until a suitable atom for bonding is found (in Figure 4.17 a,b these concentric scanning spheres are represented by red circles). If multiple atoms are found in the same range the most suitable is that which produces the most stable product, e.g. in the presence of a vinyl group with equidistant alkene atoms a bond is formed preferentially with the least hindered carbon. On bond formation the hybridization of the atom is modified appropriately and the bond length is relaxed by minimization with a high termination energy gradient of 2.062 kJ mol^{-1} $Å^{-1}$ (0.5 kcal mol^{-1} $Å^{-1}$).

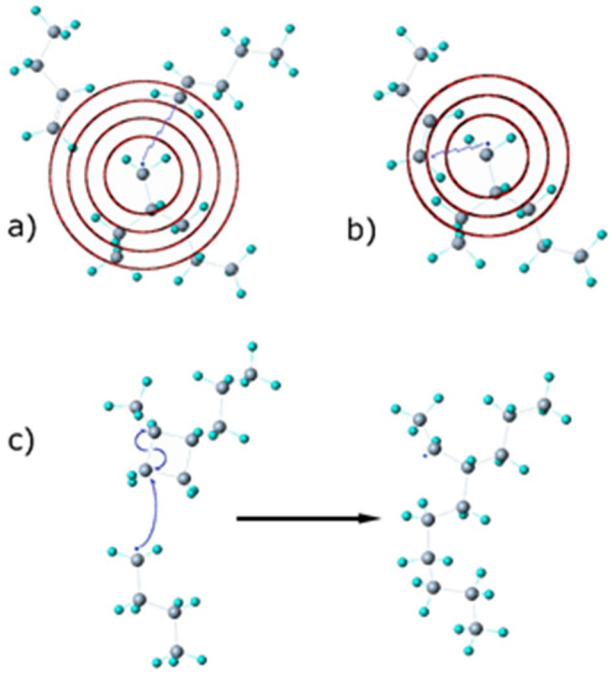


Figure 4.17. Reactions: a) by closest proximity; b) by product stability; c) with strained structures.

Parameters are included to prevent paradoxical, impossible and improbable reactions, such as bond formation between atoms which are already bound and the formation of pentavalent carbons. Other reactions are simply suppressed however, such as the formation of strained 3 and 4 membered rings. Due to the requirements for bond formation included in the algorithm these structures are prevented from occurring at rates which they might if left alone, but are permitted infrequently. However, particularly high energy strained structures formed are noted and may undergo further reactions with radicals, resulting in a ring opening process and an increase in stability (Figure 4.17c).
The process finishes when all reactions possible within the initial parameters are complete. Included in the script is a save point before polymerization begins, allowing comparative analysis of the pre- and post-polymerization systems. In analysis of the prepolymerization system a minimization of the same high energy gradient is performed, but otherwise no modifications are made on these arrangements.
Affinity analysis was performed on both the prepolymerization and polymer systems by a combination of the direct measurement of the association energy of the system and by use of docking software. The association energy, ΔE, was calculated according to equation 4.3:

$$\Delta E = E_{system} - (E_{template} + E_{system-template})$$

Equation 4.3. Association energy.

By removing the template from the system (the energy of which is given by E_{system}) and measuring the energy of both the template ($E_{template}$) and the system minus the template ($E_{system-template}$). The lowest value of ΔE thus suggests the highest binding affinity. Each template molecule was removed for analysis of $E_{template}$ and $E_{system-template}$ and replaced

sequentially for each system. The lowest value of ΔE from each was then multiplied with the Surflex-dock determined dissociation constant K_D by equation 4.4:

$$\text{Affinity score} = -\frac{\log(K_D)\Delta E}{100}$$

Equation 4.4. Affinity score.

The docking program was also used alone to predict the selectivity of the polymer and the prepolymerization system. Here the selectivity value is given by equation 4.5:

$$\text{Pre/polymer Selectivity} = \frac{\log(K_D)_{(-)}}{\log(K_D)_{(+)}}$$

Equation 4.5. Pre/polymer Selectivity.

Similarly to the empirical values, with the numerator being the affinity of the (–)-ephedrine target and the denominator being that of the (+)-ephedrine enantiomer.

4.4.3 Results and Discussion

The polymerization algorithm runs without experimenter intervention, the procedure though SA, time-evolved dynamic equilibration and polymer formation being completely automated. As a result, the size of the polymer synthesized is limited predominantly by computational power. Figure 4.18 is provided as an example of this, in which a single macromolecule is shown, built from a prepolymerization system based on composition A (minus solvent and equal to 21 template molecules), with template molecules inside. This polymer was formed in approximately 1 week using a commercially available computer, is the result of 889 new bonds and has a mass of 104 kDa, larger than some real synthesized MIP nanoparticles (Çakir et al., 2013). Figure 4.18 (left) includes an additional ephedrine molecule for comparison.

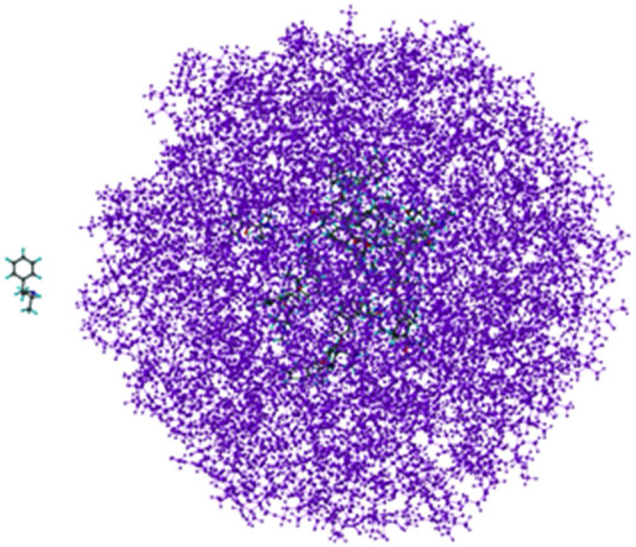


Figure 4.18. A 104 kDa MIP synthesized *in silico* by automated polymerization (right), with ephedrine template for comparison (left).

To test the validity of forming these polymers for practical purposes, comparative analysis was performed with polymerization mixtures used in the design of a (–)-ephedrine imprinted polymer. The original article (Piletsky et al., 2001) using the LeapFrog (Tripos Inc., St. Louis, MO, USA) screening method in design, an automated process in which a library of FMs are systematically placed at optimal positions around a template and ranked by their strength of interactions and suitability for application.
As discussed briefly in the introduction, the approach of analyzing the monomer-template complex has been become the dominant approach in computational design, in large part due to its simplicity. While analysis of the interactions of a small number of molecules at one time facilitated automated screening of large libraries of FMs in MM based methods, the widespread adoption of computationally expensive QM methods (DFT alone is used in approximately 56 % of all contemporary computational MIP design) has cemented this approach as the primary paradigm (Fu et al. ,2003; Baggiani et al., 2005 and Dong et al., 2005). Some doubt has been cast on the validity of this approach however by research such as that by Sobiech et al., in which observations of the monomer-template complex were followed by analysis of a more developed system of monomer, cross-linker and template (Sobiech et al., 2014). In comparison with empirical data the systems including cross-linkers predicted the same relative affinities as that of the experimental data, while the models using monomer-template complexes alone showed little correlation. Attempts to further develop alternative models for design involving the formation of small oligomers in the past have had significant limitations, being largely formed by manually connecting random combinations of monomers or a small number of molecules already complexed to the template. One exception to this is demonstrated in the research of Schauperl and Lewis, in which a polymer was grown by an evolutionary process around a number of templates using their ZEBEDDE zeolite modeling program (Schauperl and Lewis, 2015). While this research provides an excellent example of sophisticated design procedures be implemented, the method used still suffered from some problems of representation of real polymer synthesis.
The use of MD has been less widely used in MIP design, despite the excellent theoretical research built on its application (Golker and Nicholls, 2016, Golker et al., 2014, Dourado

et al, 2011; Nicholls et al., 2009, Karlsson et al, 2009; Pavel and Lagowski, 2005, Molinelli, et al., 2005). MD permits analysis of the whole prepolymerization mixture and its evolution through time to a point of equilibrium, providing the most accurate model of the system currently available (Ansell and Luah, 2005, Andersson et al., 1999, Katz and Davis, 1999, Olsson et al.,2012, Nicholls et al., 2009, Nicholls et al., 2009). Assumptions required in analysis of the monomer-template complex relating to its formation in solution and lack of response to other system components are not necessary in MD. However, the assumption that the interactions of the template in the prepolymerization mixture of monomers will be proportional to the interactions with the imprinted polymer remains.
MD was used in this research to approximate a realistic arrangement of molecules in the mixture before inducing polymerization. Analysis was then performed by a combination of docking and direct association energy assessment with both the prepolymerization mixture and the final polymer, and compared with empirical data and LeapFrog (Tripos Inc., St. Louis, MO, USA) based monomer rankings. Three different FMs were examined, first for the predicted affinity of each approach. The results are shown Table 4.11.

Table 4.11. Empirical affinity results and theoretical predictions with three MIP composition.

ID	Functional monomer	Empirical imprinting factor (Piletsky et al., 2001)	Empirical capacity factor (Piletsky et al., 2001)	LeapFrog binding score [kcal mol $^{-1}$] (Piletsky et al., 2001)	Pre-polymer affinity	Polymer affinity
A	MAA	2.87	9.46	-14.62	-4.093	-4.530
B	HEM	2.67	0.80	-15.72	-3.162	-3.165
C	2VP	1.00	0.10	-1.82	-3.382	-2.316

Two sets of empirical data are included to demonstrate the affinity of the MIPs for the ephedrine target, an imprinted factor and a capacity factor. Both are included to demonstrate the trend in affinity displayed with the highest binding being associated with the MAA functionalized MIPs and the lowest with the 2VP. Inclusion of both sets of data is intended to provide some element of consilience, the mesoscopic polymer properties which may positively or negatively affect each composition being cancelled to some extent by the different measurements. The theoretical results in Table 4.11 are all presented in terms of relative energetic stability, and a smaller (more negative) value thus represents greater binding. The LeapFrog screening results from the original article (Piletsky et al., 2001) are presented in the fifth column, and demonstrate the utility of the approach in eliminating the weakly interacting functional monomer (2VP). This screening however would lead the observer to incorrectly predict that composition B would result in the strongest binding MIP.
The sixth column ('Pre-polymer affinity') shows the interactions observed in the prepolymerization equilibrated system. Here the most effect system is correctly identified

as that with the MAA monomers, but the other two compositions are incorrectly ordered in terms or relative affinity displayed. The values obtained for each polymer formed from each of these systems are then shown in the final column, and can be observed to follow the results obtained in the empirical investigation.

Table 4.12. Selectivity values for the three MIP compositions, with results of pre- and post-polymerization analysis by docking.

Functional monomer	Empirical selectivity $k`_{(-)}/k`_{(+)}$ (Piletsky et al., 2001)	Pre-polymer selectivity	Polymer selectivity
MAA	1.34	1.13	1.54
HEM	1.42	1.25	1.57
2VP	1.00	1.19	1.04

Docking software allow easy analysis of binding site selectivity, and so this was performed on the prepolymerization mixtures and equivalent polymers using the template (–)-ephedrine and its enantiomer (+)-ephedrine. The results are shown in Table 4.12. In the empirical analysis the MIP exhibiting the greatest selectivity is the HEM polymer, while the 2VP MIP shows no selectivity. In the prepolymerization system the most successful composition is again identified, but the differences between the values for each MIP are relatively small. The 2VP polymer is also predicted to displayed notably more selectivity than was observed, outperforming the underestimated MAA MIP. The polymer models here perform well again, with the ordering being correctly predicted. The 2VP system also displays almost no selectivity, despite the high values associated with the other two compositions.

4.4.4 Conclusions

An algorithm which mimics the radical polymerization has been produced for application in the design of molecularly imprinted polymers. Written into this algorithm is the capacity for SA and equilibrating MD, meaning that an entirely automated process can bring a random arrangement of molecules through to a state likely to be an appropriate model of the prepolymerization mixture and polymer synthesis performed without experimenter intervention. In investigating the design of a synthetic ephedrine receptor this approach was observed to produce values for both the relative affinity of the target for the receptors and the enantioselectivity associated with each. The method described is believed to produce the most accurate atomistic models of molecularly imprinted polymers thus produced. The project was performed by a close cooperation with the Leicester Biotechnology Group.

Chapter 5

Molecular characterization of HIV-1 Nef and ACOT8 interaction: insights from *in silico* structural predictions

The results described in this chapter have been published in Serena et.al., 2016 paper.

5.1 Introduction

Nef. HIV-1 Negative factor (Nef) is a protein essential for the metabolism of the HIV-1 virus and it takes part in several mechanisms modulating the virus infectious phase (Kestier et.al.,1991). Furthermore, in humans, several long-term survivors of HIV infection carry HIV with deletions in Nef or with a high frequency of defective Nef alleles (Salvi et.al., 1998; Deacon et.al., 1995 and Kirchhoff et.al., 1995). In tissue cultures, several functions of Nef have been described. Among them, five are the most important. In the first place, Nef enhances viral infectivity and replication in primary cells (Spina et.al., 1994 and Miller et.al., 1994). In the second place, the HIV-1 accessory protein alters the state of T-cell activation and macrophage signal transduction pathways (Swingler et.al., 1999; Iafrate et.al., 1997; Graziani et.al., 1996 and Skowronski et. al., 1993). Third, Nef inhibits the immunoglobulin class switching (Qiao et.al., 2006). Forth, it reduces the cell surface expression of CD4. Indeed, the internalization and degradation of cell surface CD4 by Nef augments the infectivity of the released virion particles because CD4 interferes with incorporation of the HIV envelope protein into the virus particle. Eventually, Nef downregulates the cell surface expression of MHC-I molecules. This fact allows Nef both to escape the host immune response (Cohen et.al., 1999 and Zipeto and Berreta, 2012) and associates with several components of the endocytic pathways (Chaudhuri et.al., 2007). Nef, by using its N-terminal myristoylation, is able to bind the cell membrane. Indeed for this reasons it is also known as raft-associated protein (Bentham et.al.,2006 and Wang et.al., 2000). Myristoylated Nef can adopt several quaternary structures as monomers, dimers and trimers and it may associate with other proteins (Arold et.al., 2000 and Kienzle et.al., 1993). It is worth mentioning that the only myristoylation of Nef is not enough to bind the membrane so more complex associations are required to allow its migration and binding to the membrane. (Bentham et.al., 2006).

Acyl-coenzyme A thioesterase 8. Human thioesterase 8 (ACOT8) is a protein that interacts with Nef (Liu et.al.,1997 and Watanabe et.al.,1997). ACOT8 is a peroxisomal enzyme and it is important for the lipid metabolism. The human ACOT8 gene is located on chromosome 20q13.12 and the latter codes for a protein of 319 aminoacids and a molecular mass of 35 kDa (Liu et.al., 1997 and Jones et.al., 1999). This protein is in the peroxisomes because of a serine-lysine-leucine (SKL) peroxisomal targeting signal (Liu et.al.,1997; Jones et.al.,1999 and Miura et.al.,1992). It has been reported that murine ACOT8 is inhibited by Coenzyme A (CoASH) (Hunt et.al., 2002), diversely from the Type-I ACOTs. Two factors such as the sensitivity to CoASH and the very broad substrate specificity indicate that ACOT8 has a specific function in tuning the intra-peroxisomal acyl-CoA/CoASH level with the purpose of optimize the fatty acids flux

through the β-oxidation system. ACOT8 compared with the peroxisomal Type-I ACOTs, exhibits a large tissue expression range both in mice and humans (Watanabe et.al., 1997 and Hunt et.al. ,2002). Anyhow, so far, the ACOT8 function in lipid metabolism have not been elucidated. So far, ACOT8 structure has not been crystallized yet. However, Li and co-workers (Li et.al.,2000) solved the 3D structure of the E. coli thioesterase II by X-ray at 1.9 Å resolution. The two proteins share ca. 41% of SI. While E. coli thioesterase II exists as a tetramer, the ACOT8 is present both in dimeric and tetrameric forms (Cohen et.al., 2000). Yeast two-hybrid studies have shown that HIV-1 Nef directly interacts with ACOT8 (Liu et.al., 1997 and Watanabe et.al., 1997). HIV-1 Nef-LAI region spanning residues 108-124 (in particular Asp^{108}, Leu^{112}, Phe^{121}, Pro^{122}, Asp^{123}) have been shown as crucial for ACOT8 association (Cohen et.al., 2000 and Liu et.al., 2000). Moreover the expression of ACOT8 facilitates the relocalization of Nef to peroxisomes in 3T3 cells (Cohen et.al., 2000). Nef/ACOT8 colocalization in peroxisomes requires the C-terminal peroxisomal targeting sequence of ACOT8.
Over the time, several speculations have been advanced to understand why HIV-1 Nef associates with ACOT8. The preferred substrates of ACOT8 are myristoyl-CoA and palmitoyl-CoA (Liu et.al., 1997); indeed ACOT8 could participate in the regulation of lipid modifications of proteins, which are relevant for their membrane binding and receptor internalization (Liu et.al., 2000). Previous studies described that palmitoylation may affect the velocity of endocytosis of molecules at the plasma membrane level (Alvarez et.al., 1990 and Eason et.al., 1994). Therefore, ACOT8 could affect the acylation of these macromolecules by tuning the intracellular level of acyl-CoA. Moreover, lipid rafts are preferential sites for HIV-1 viral particles budding (Waheed and Freed, 2009) and it has been reported that hydrolysis of long chain fatty acyl-CoA is required for correct budding of Coat Protein Complex I (COP-I) coated vesicles (Rothman, 1994), whose β-COP subunit is a Nef interacting target (Benichou et.al.,1994). In the absence of acyl-CoA, buds accumulate and coated vesicles fail to pinch off. Since ACOT8 hydrolyses a broad range of acyl-CoA and Nef may enhance its enzymatic activity in a dose dependent manner (Watanabe et.al., 1997), it is likely that the Nef-mediated ACOT8 recruitment may be involved in intracellular transport pathways modification. So far, research studies have been focused on determining the Nef regions involved in the interaction with ACOT8 and the lack of structural information for ACOT8 limits the full comprehension of the biological role of the Nef/ACOT8 association relevant to HIV-1 infection.

5.2 Aims of the research described in this Chapter

- To model the human ACOT8 structure
- To predict the ACOT8 residues most probably involved in the interaction with Nef
- To generate a Knowledge-based Nef/ACOT8 complex model able to satisfy all the findings of this study and of previous works.
- To gain insights into the difference between Nef-LAI and Nef-SF2 allelic variants.

5.3 Materials and Methods

ACOT8 homology modeling. To generate the ACOT8 model, protein sequences of all members of the ACOT8 family were taken from the Uniprot database. A MSA was built using the PROMALS program (Pei et.al., 2007 and Söding, 2005). The sequences were used for the creation of a HMM profile of the ACOT8 family by using the HHsearch program (Söding, 2005). All alignments were displayed using the ESPript 3.0 webserver (Robert and Gouet, 2014). The MSA obtained was used as reference for the structural modeling of ACOT8. The 3D structural predictions for ACOT8 was built by homology modeling using the *E. coli* thioesterase (PDB code: 1C8U) as template. Then, the dimeric functional configuration was built and energy minimized using the MODELLER9v3 program (Eswar et.al., 2008).

Prediction Nef/ACOT8 interface and Dockings: CPORT-HADDOCK. To characterize the Nef/ACOT8 interaction, the 3D structure of Nef-LAI (PDB: 1AVV) was exploited. Since that the region spanning aminoacids from 148–178 is missing, this segment was completed by using the structure of Nef-LW123 (PDB: 2NEF). However, in the two structures, the experimental binding regions were identical and the added loop was far from that binding region. The Nef/ACOT8 electrostatic potentials were calculated using the APBS program (Baker et.al., 2001). A bioinformatic predictor called CPORT, was used to predict protein-protein interface residues (de Vries and Bonvin et.al., 2011). The Nef/ACOT8 docking model was then performed using the HADDOCK (High Ambiguity Driven biomolecular DOCKing) version 2.1 program (de Vries et.al., 2007 and Dominguez et.al., 2003), taking into account our structural bioinformatics predictions. These predictions were funneled as AIRs in the docking process. The docking experiment was carried out by using default parameters through the web-server version of HADDOCK (Van Zundert et.al, 2016 and Dominguez et.al., 2003). All calculations were carried out by CNS1.2 (Brünger et.al., 1998). Non-bonded interactions were calculated with the OPLS force field using a cut-off of 8.5 A (Jorgensen and Tirado-Rives, 1988). The electrostatic potential was calculated using a shift function, while a switching function (between 6.5 and 8.5 A) was used to define the Van der Waals potential. The AIRs to guide the docking were defined as follows. Asp^{108}, Leu^{112}, Phe^{121}, Pro^{122} and Asp^{123} (Cohen et.al., 2000 and Liu et.al., 2000) in the Nef protein were considered as active residues, while for ACOT8 the active residues were Arg^{43}, Arg^{45}, His^{46}, Lys^{52}, Arg^{53}, Arg^{86} and Lys^{91}. They were chosen considering both the charge complementarity between Nef and ACOT8, and the CPORT (de Vries and Bonvin et.al.,2011) predictions. Passive residues were not explicitly selected but they were automatically defined by the web server as those in proximity of the active ones. Docking decoys were displayed using the CHIMERA program (Pettersen et.al., 2004).

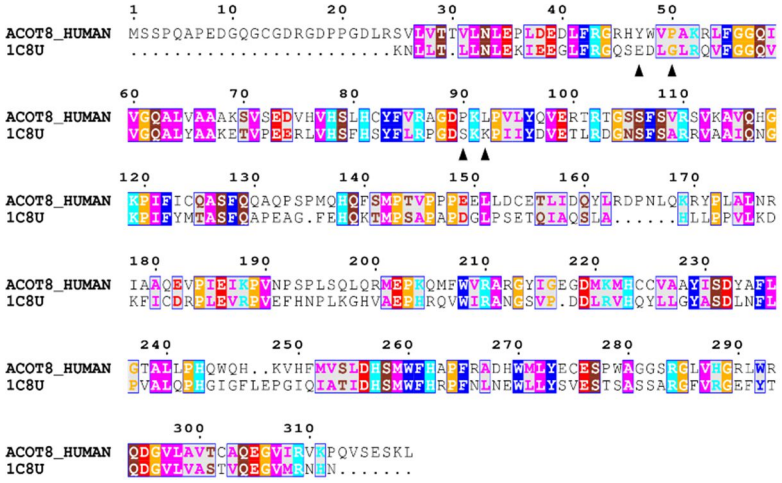
Knowledge-based Nef/ACOT8 model. To guide the second docking experiments, the residues considered as active for Nef were the same as the first docking and no passive amino acids were defined. For ACOT8, only Lys^{91} was chosen as active residue and Arg^{86}, Ala^{87}, Gly^{88}, Asp^{89}, Pro^{90}, Leu^{92} and Pro^{93} were defined as passive residues.

5.4 Results

5.4.1 ACOT8 homology modeling

For the modeling procedure, the pairwise sequence alignment between ACOT8 and the E. coli thioesterase (Fig. 5.1a) was extracted from the Hidden Markov profiles generated using the HHpred program (Söding et.al., 2005). A set of models comprising 100 decoys was generated by MODELLER then the final working structural model was selected among these structural predictions by using the MODELLER objective functional scores. All the generated 3D models of ACOT8 showed high structural similarity with the currently available experimental geometries, indicating that both the secondary structures elements (12-stranded antiparallel β -sheets) and the typical tertiary fold (the double hot dog) (Li et.al., 2000) were preserved in the model. The interface between the two subunits was totally buried in the protein, involving the central fragments of the six central β-sheets from the two monomers to form several stabilizing interactions (Fig. 5.1b). The contact surface of ACOT8 was structurally similar to that of E. coli enzyme, i.e. 1980 and 2142 A^2 per monomer for the target and the template, respectively.

a)



b)

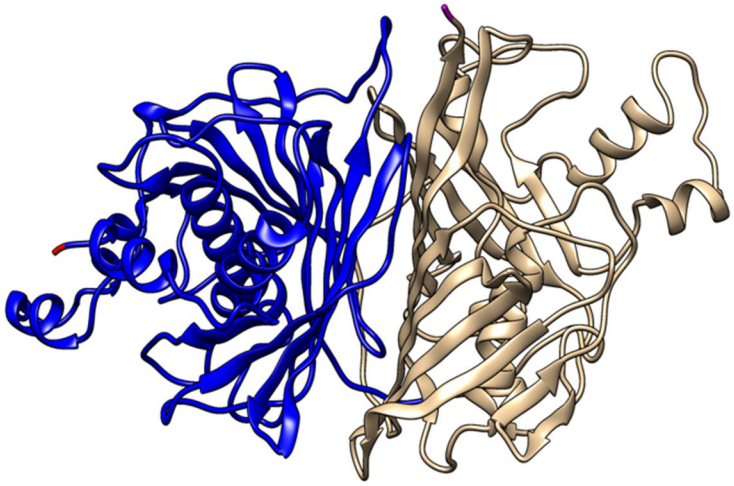
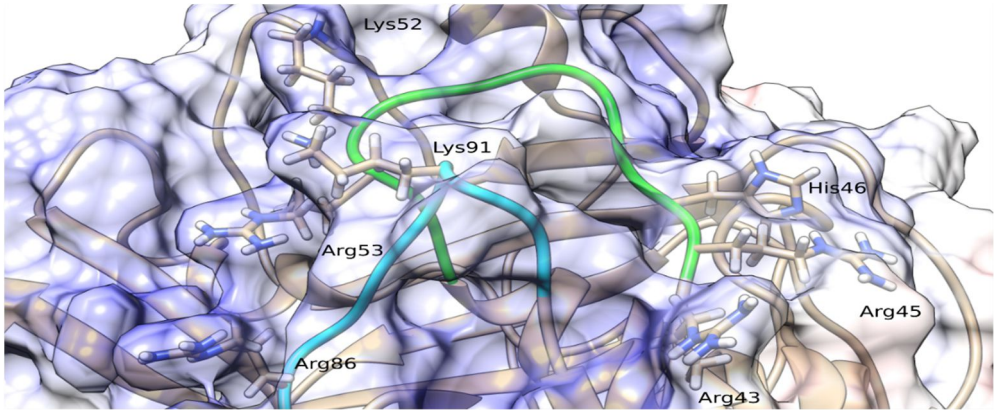


Figure 5.1. (a) Sequence alignment between Acyl-coenzyme A thioesterase 8 (ACOT8_HUMAN) and E. coli thioesterase II (1C8U). ▲ indicates aromatic and aliphatic residues (Tyr^{47}, Pr^{o50}, Pro^{90} and Leu^{92}) of ACOT8, which are replaced by non-hydrophobic residues in E. coli isoenzyme (Glu, Gly, Ser and Lys respectively). The residues are coloured according to their physico-chemical properties (HKR in cyan, DE in red, STNQ in maroon, AVLIM in pink, FYW in blue, PG in orange and C in green). Reprinted from Serena et.al., 2016 article. **(b) The modelled ACOT8 dimer (chain A in blue and chain B in brownish).** The homology model of ACOT8 was built based on the X-ray structure of the E. coli thioesterase. Reprinted from Serena et.al., 2016.

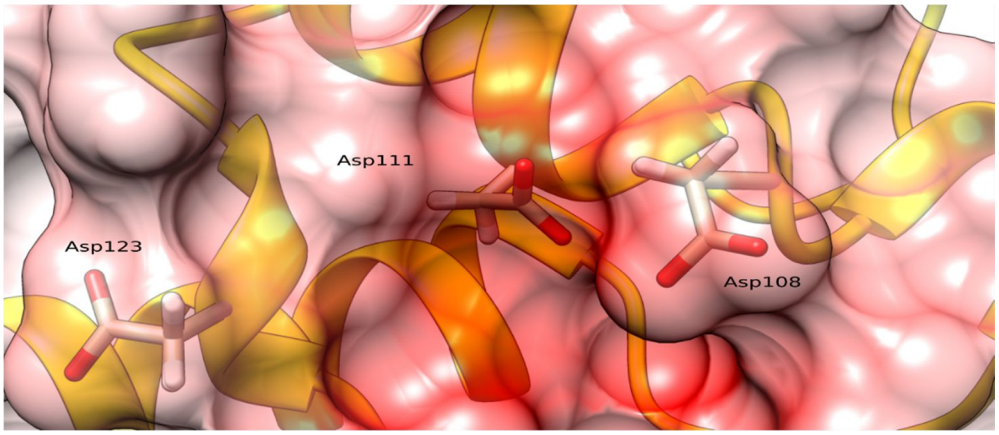
5.4.2 Initial model of Nef/ACOT8 interaction

The electrostatic potential on Nef/ACOT8 surfaces (Fig. 5.2a, b) show a high electrostatic complementarity, i.e. a negatively charged surface on Nef-LAI and a positively charged one on ACOT8. Moreover, the latter corresponded to one of the ACOT8 highest probability interaction regions detected on the molecule through the CPORT (de Vries and Bonvin, 2011), which offers a prediction score for each residue of the ACOT8 (Fig. 5.2c).

a)


Lys52
Lys91
His46
Arg53
Arg45
Arg86
Arg43

b)


Asp111
Asp123
Asp108

c)

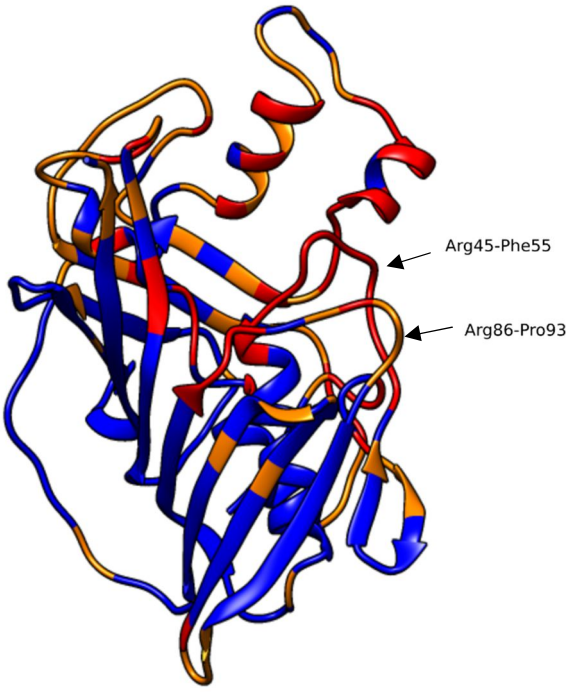

Arg45-Phe55
Arg86-Pro93

Figure 5.2. Electrostatic potentials on the ACOT8 (a) and the Nef (b) surfaces. The Nef/ACOT8 electrostatic complementarity can be vividly visualized by the plots of the electrostatic potential, calculated solving the Poison-Boltzmann equation. The contact surfaces of Nef and ACOT8 are clearly complementary: while the surface of Nef is highly negative (red), the contact surface of ACOT8 is highly positively charged (blue) (reprinted from Serena et.al., 2016 article). **(c) CPORT prediction**. The ACOT8 residues with a high prediction score are likely supposed to be involved in the interaction with Nef. In the figure, red/orange colour means a high prediction score, whereas blue colour means a low prediction score. Reprinted from Serena et.al., 2016.

The aminoacids with a high-predicted score can presumably participate in the association with Nef. These computational results and the available experimental data on Nef (Cohen et.al., 2000 and Liu et.al., 2000), were combined together to drive the docking phase using HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003), to gain a deeper understanding of the Nef/ACOT8 interaction (Fig. 5.3).

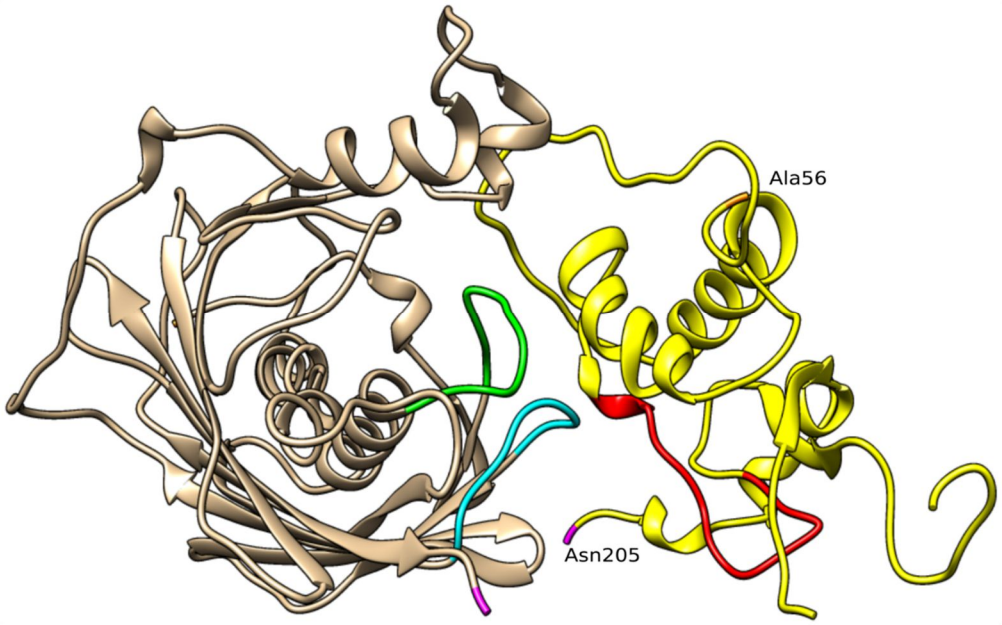


Figure 5.3. Docking model. Nef (yellow) and ACOT8 (brownish) docking model. The ACOT8 Arg^{45}-Phe^{55} and Arg^{86}-Pro^{93} regions are highlighted in green and in cyan respectively. For both proteins, N- and C-terminals are coloured in orange and magenta respectively. Nef amino acids, starting from Asp^{123}, involved in proteasome degradation are highlighted in red. Reprinted from Serena et.al., 2016.

HADDOCK (Van Zundert et.al., 2016 and Dominguez et.al., 2003) successfully generated all the structures. The structures were clustered by using default parameters and five clusters were produced, among them, only two were sufficient populated.
The most populated cluster (111/200 decoys) of decoys contained the best scoring models. In particular, the root-mean-square-deviation (RMSD) of the backbone atoms from the overall lowest-energy structure was 0.64 ± 0.40 Å. The second most populated comprising just 53/200 decoys, and while the HADDOCK score was about −222.6 for the first cluster, in the second it was −96.7. Thus, the representative structure of the most

populated cluster was considered in order to gain structural information on the aminoacids present in the Nef/ACOT8 interface. In this cluster, ionic interactions were found between Lys^{52} of ACOT8 and Asp^{108} and Asp^{111} of Nef-LAI, but also between Lys^{91} of ACOT8 and Asp^{123} of Nef. Beyond these strong electrostatic interactions, 11 hydrogen-bond interactions were detected between Nef and ACOT8. Finally, hydrophobic contacts occured between three residues of Nef, i.e. Leu^{112}, Phe^{121} and Pro^{122} and three residues of ACOT8, i.e. Tyr^{47}, Pro^{50} and Pro^{90}.

5.4.3 Knowledge-based Nef/ACOT8 complex model

Our initial structural bioinformatics prediction discussed in the previous 5.4.2 Initial model of Nef/ACOT8 interaction section enabled the identification of the ACOT8 residues putatively involved in the interaction with Nef. In particular, they were found in the Arg^{45}-Phe^{55} and Arg^{86}-Pro^{93} ACOT8 regions. Our collaborators confirmed our predictions by *in vitro* assays through the development of ACOT8 deletion mutants. Furthermore, immunofluorescence and co-immunoprecipitation studies highlighted that the ACOT8 K91S mutation is sufficient to disrupt the interaction with Nef. Moreover, they also experimentally confirmed that the two ACOT8 binding regions i.e. Arg^{45}-Phe^{55} and Arg^{86}-Pro^{93} region were indispensable in binding and stabilization with Nef. These findings were used to recreate a state of the art structural model able to explain all the experimental data, that we have named in the original paper (Serena et al., 2016), as *"Knowledge-based Nef/ACOT8 model"*. To do this, a new virtual protein-protein docking between the structural model of Nef-LAI and ACOT8 was performed. The docking procedure was drive by using in *vitro* experimental data. The docking complexes resulting from this computational experiment were clustered in three different clusters. The best score model of the most populated cluster (165/200 decoys) was examined by visual inspection. Several pairs of residues were involved in strong intramolecular salt bridges. Indeed, while ACOT8 shows a positively charged interaction surface, formed principally by Arg^{43}, Arg^{45}, His^{46}, Lys^{52}, Arg^{53}, Arg^{86} and Lys91, the Nef negatively charged surface is formed by Asp^{108}, Asp^{111}, Asp^{123}. Four strong ionic interactions were found. In particular, between ACOT8 Lys^{91} and Nef Asp^{108} and Asp^{111}, and also between ACOT8 Arg^{53} and Nef Asp^{108}, and finally between the ACOT8 Arg^{43} and Nef Asp^{123}. We saw that Lys^{52} did not bind any Nef aminoacids, as observed by *in vitro* experiments. However, the region spanning aminoacids from Arg^{45} to Phe^{55} appeared to be involved in the complex stabilization. Several intermolecular hydrophobic contacts were also formed between three residues of Nef, i.e. Leu^{112}, Phe^{121} and Pro^{122} and five residues of ACOT8 i.e. Tyr^{47}, Pro^{50}, Pro90, Leu^{92} and Pro^{93}, allowing a further stabilization of the complex (Fig. 5.4). This hypothesis could be supported by the observation that some of these hydrophobic amino acids, i.e. Tyr^{47}, Pro^{50}, Pro^{90}, Leu^{92}, are lacking in the *E. coli* thioesterase and this may be the reason why the interaction with Nef does not take place, as previously observed by Cohen and colleagues (Cohen et.al., 2000). In fact, in the E. coli thioesterase, these residues are substituted by non-hydrophobic residues (Glu, Gly, Ser and Lys respectively), completely changing the physico-chemical properties of the interaction surface (Fig. 5.4).

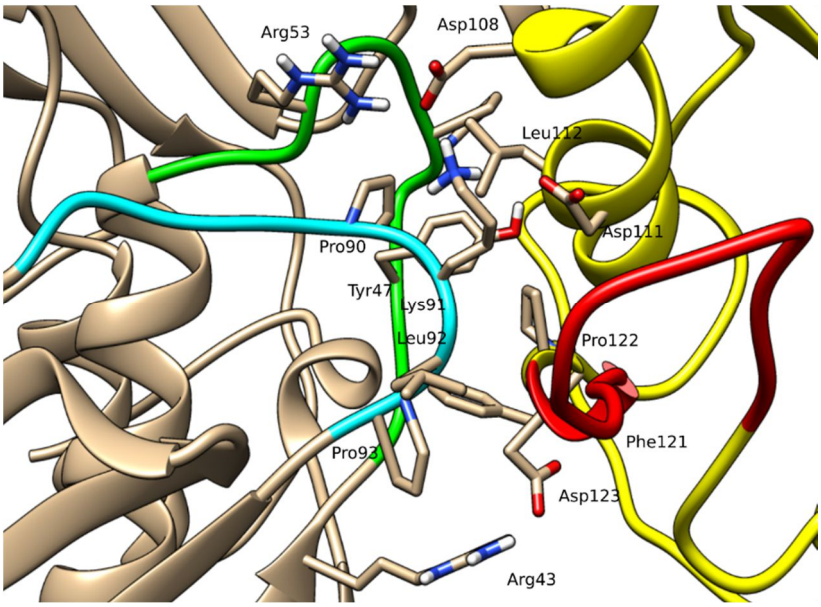
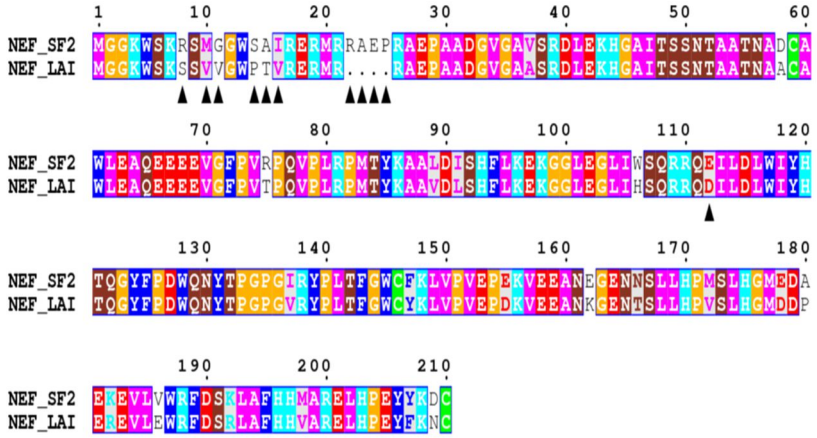


Figure 5.4. Detail of the knowledge guided Nef/ACOT8 complex. Residues involved in the Nef/ACOT8 interaction, characterized in this work for ACOT8 and by Liu and Cohen (Liu et. al., 2000 and Cohen et.al., 2000) for Nef, are shown. The colour code is the same as in Fig. 5.3. Reprinted from Serena et.al., 2016.

5.4.4 Comparison between Nef-LAI and Nef-SF2

Our collaborators investigated putative difference in ACOT8 binding between the two most common Nef variants i.e. Nef-LAI and Nef-SF2. This study was performed by coimmunoprecipitation analysis and they showed that ACOT8 associated only with Nef-LAI variant, but not with the Nef variant from HIV-1-SF2. In order to gain a deeper understanding of the putative differences between Nef-LAI and Nef-SF2 variants, structural bioinformatics studies were performed. Nef-LAI and Nef-SF2 differ by 30 aminoacidic residues. Looking into more details, the sequence alignment (Fig. 5.5a), and the Nef-LAI region spanning aminoacids from 108 to 124, it was appreciated that Asp^{108} is swapped in glutamic acid in Nef-SF2 (residue 112 in Nef-SF2 homologous). Cohen and collaborators demonstrated that this single mutation is responsible for the lack of the key ionic bond with the human protein (Cohen et.al., 2000). In the Nef-SF2 structure the longer side-chain of the glutamic acid could form an intramolecular ionic bond with the closest Arg^{109} (following the Nef-SF2 numbering system) (Fig. 5.5b). Hence, the aminoacids would not be able to interact with ACOT8, weakening an important salt bridge network and destabilizing the Nef/ACOT8 interaction. It has been also verified that the N-terminal region of Nef-LAI could participate either in the interaction or in the stabilization of the Nef/ACOT8 complex (Cohen et.al., 2000). Unluckily, this portion was not solved for the Nef-LAI crystal structure, and thus we did not include it in the model.

a)



b)

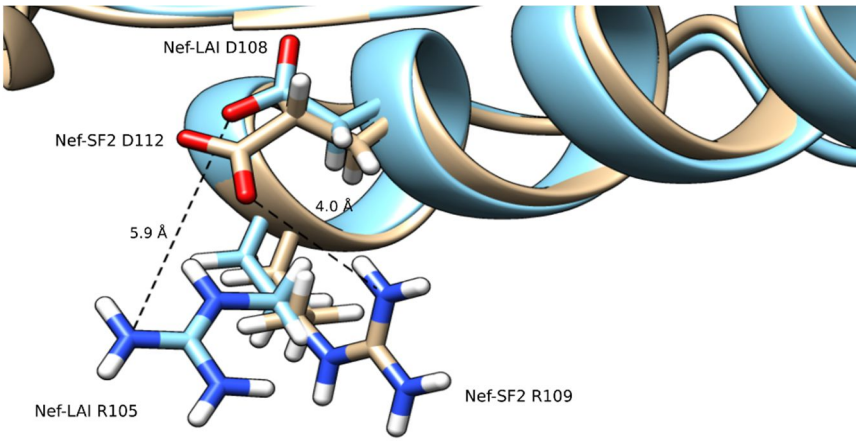


Figure 5.5. Sequence alignment between Nef-SF2 and Nef-LAI. (a) ▲ indicates some residues differences between the two alleles in the N-terminal region and in Glu^{112} corresponding to Asp^{108} in Nef-LAI. Rreprinted from article Serena et.al., 2016. **(b)** Superimposition of Nef-LAI structure (in cyan) and Nef-SF2 structure (in brownish) shows that, while Nef-SF2 Glu^{112} might interact with the Nef-SF2 Arg^{109} due its longer side-chain, Nef-LAI Asp^{108} (shorter side-chain) could not be involved in an intramolecular salt bridge with Nef-LAI Arg^{105}, remaining free to interact with ACOT8. Reprinted from article Serena et.al., 2016.

5.5 Discussion and Conclusions

In literature it has been reported that the only known viral protein interacting with ACOT8 is HIV-1 Nef. However, so far, the ACOT8 structure and the regions interacting with Nef have not been characterized yet. Indeed, for many years the role of the Nef/ACOT8 association remained unclear.

Our work has the main goal to identify, for the first time, the key ACOT8 aminoacids involved in the interaction with Nef and to gain insights into the biological role of their interaction.

As first step the ACOT8 regions most likely involved in the interaction with Nef were successfully predicted by state of the art bioinformatics tools. Subsequently, several mutants were designed based on our bioinformatics predictions. The latter were successfully confirmed by our collaborators and it was demonstrated that ACOT8 Lys^{91} plays a key role in the interaction with Nef. Moreover, both ACOT8 Arg^{45}-Phe^{55} and Arg^{86}-Pro^{93} regions were identified as determinant for Nef association, indicating that the interaction involves a larger portion on ACOT8 surface. In this work, we have also computationally and experimentally characterized the differences between the two main Nef allelic variants. In particular, by western blot analysis we noticed that Nef-LAI associated with ACOT8. Indeed, the more ACOT8 is expressed, the more Nef-LAI is detected. On the other hand, no effect was appreciated when ACOT8 was cotransfected with Nef-SF2. Thus, we hypothesized that the ACOT8 association to Nef could have a stabilizing role for the Nef expression. Interestingly, it has been reported that, in the first step of HIV-1 replication, proteasome can interact with Nef (Schwartz et.al., 1998). Indeed, it has been demonstrated that the Nef 123 to 152 region (Nef 123–152) is one of the immunogenic domains involved in proteasomal digestion and that the region comprising amino acids 128 to 135 is one of the principal cleavage products (Lucchiari-Hartz et.al., 2000). The Nef 123–152 segment is highly conserved among different HIV-1 Nef variants (da Silva et al., 1998) (as indicated in Fig. 5.5a) but, since this region overlaps with the Nef/ACOT8 interaction surface, we guess that when Nef binds ACOT8, the proteasome could not interact, probably due to steric hindrance. Our data support the hypothesis that the Nef association to ACOT8 may improve the Nef stability, probably preventing its proteasome-mediated degradation. To conclude, this experimental/computational combined study, through the identification of the ACOT8 regions most likely involved in the interaction with Nef elucidates the association between HIV-1 Nef and human ACOT8. Our data paved the way for further studies aimed at broadening the comprehension of the biological role of their interaction.

In the two next chapters, we will show two examples of multiscale computational biophysics applications on two different targets.

Chapter 6

Molecular dynamics simulations of F152I-Mi and I244T-Mi variants associated with PH1 and implications in their pathogenicity

The results described in this chapter have been published in Dindo et al., 2016 paper.

6.1 Introduction

Human liver peroxisomal alanine glyoxylate aminotransferase (AGT) (EC 2.6.1.44) catalyses the transamination of L-alanine and glyoxylate to pyruvate and glycine, respectively, by a classical ping-pong mechanism.
The process involves two steps. In the first half-reaction the AGT in the pyridoxal 50-phosphate (PLP) form (AGT-PLP) reacts with L-alanine generating pyruvate and the AGT in the pyridoxamine 50-phosphate form (AGT-PMP). Afterward, in the second half reaction, AGT-PMP binds glyoxylate and transforms it to glycine reforming AGT-PLP (Cellini et al., 2007). Human AGT is a dimer protein and belongs to the Fold Type I class of the PLP-enzyme family (Zhang et al., 2003).
The typical 3D structure of AGT consists of an N-terminal extension (residues 1–21) that wraps over the surface of the other subunit, a large domain (residues 22–282) containing most of the active site residues and the dimerization interface, and a small domain (residues 283–392) containing a peculiar KKL type 1 peroxisomal targeting sequence (PTS1) at the C-terminus (Zhang et al., 2003). Mutations in the AGXT gene encoding human AGT provoke a disease called Primary Hyperoxaluria Type I (PH1) (OMIM 259900) (Danpure, 1986).
Primary Hyperoxalurias are metabolic recessive disorders caused by inborn errors in the metabolism of glyoxylate and oxalate. There are three types of Primary Hyperoxalurias, but the PH1 form is the most frequent (80% of the patients) and the most aggressive.
PH1 has an estimated predominance ranging from 1 to 3 per million population being more frequent in countries where consanguineous marriages are common (Salido et al., 2012 and Williams et al., 2009).
In humans, two polymorphic forms of the AGXT gene encode AGT protein. One form is the major allele that encodes AGT-Ma and it is the most usual form. The other form is the minor allele that encodes AGT-Mi. AGT-Mi diverges from the AGT-Ma for the presence of a 74- bp duplication in intron 1 and of two mutations causing the Pro11-to-Leu and Ile340-to-Met residues substitutions (Purdue et al., 1991).
The P11L substitution provokes a structural motif of binding to the mitochondrial import receptor TOM 20, and this might partly elucidate why AGT-Ma is 100% peroxisomal, while AGT-Mi is 95% peroxisomal and 5% mitochondrial. Furthermore, AGT-Mi differs from AGT-Ma because it has a reduced (~70%) catalytic activity and, principally, a reduced dimeric stability under chemical stress (Cellini et al., 2010 and Pey et al., 2011). Nevertheless, since the true value of the dimer-monomer equilibrium dissociation constant ($K_{d(dim\text{-}mon)}$) for both AGT-Ma and AGT-Mi is only assumed to be lower than

0.3 mM (Cellini et al., 2010), it is not possible to establish if the mutations P11L and/or I340M alter the dimer-monomer dissociation under native conditions.
The AGT-Mi polymorphism is not pathogenic per se, but many mutations are pathogenic only when they functionally synergize with the minor allele. Some of them, like G170R, F152I, I224T, G47R and G41R cause a complete or partial aberrant targeting of AGT to mitochondria, thus resulting in a hampering of the metabolic function. Fargue et al. (Fargue et.al., 2013) proposed the hypothesis that all PH1-causing mutations segregating with AGT-Mi synergize with the P11L mutation unmasking the cryptic mitochondrial targeting sequence, and, thus, provoking the peroxisome-to-mitochondrion mistargeting. On the other hand, it worth mentioning that, when transiently expressed in CHO cells, the G161 variants co-segregating with AGT-Mi were detected as cytosolic aggregates (Oppici et al., 2013). Moreover, it was described that (i) the G41RMi shows a complex picture, some mithocondrial labelling and some peroxisomal in the shape of intraperoxisomal aggregates (Cellini et al., 2010), and (ii) G47R-Mi shows a peroxisomal and mitochondrial localization (Montioli et al., 2015). Finally, it is worth mentioning that a clear depiction of the structural basis of the proposed synergic interaction of P11L with each mutation causing the mistargeting does not exist.

6.2 Aim of the research described in this Chapter

- to predict, by using MD simulations, the structural effects caused by the mutations F152I and I244T on AGT-Mi, associated with Primary Hyperoxaluria type 1, providing a possible interpretation and explanation for their interference with the dimerization process.

6.3 Material and Methods

MD simulations. All MD simulations were carried out by using the software GROMACS 4.5.5 (Pronk et al., 2013). GROMOS 53A6 united atom force field (Oostenbrink et al., 2004) and the explicit simple point charge (spc/e) water model (Berendsen et al., 1981) were used for the simulations. The dimeric structure of human AGT was obtained by means of the PISA web server starting from the available coordinate file of the monomer (PDB id: 1H0C). In the original structure, loop residues 120-123 are missing, so the latter were modelled by a loop modeling script of the MODELLER software (Webb and Sali, 2014). We modelled AGT-Mi by introducing the P11L and I340M mutations in the AGT protein using the Chimera software (Pettersen et al., 2004). The latter was also used to introduce the F152I and I244T mutations on AGT-Mi. The molecular topology and atoms parameters for PLP covalently bound to lysine 209 were taken from ATB (Kass et al., 2014 and Malde et al., 2011). The systems were simulated in a triclinic water box, and the minimum distance between any atom of the protein and the box wall was set up to 1.0 nm. The systems were neutralized by adding Na^+ and Cl^- ions using the genion program of the GROMACS 4.5.5 package (Pronk et al., 2013). Each system was energy minimized in four phases. In the first two phases, 800 steps of steepest descent algorithm followed by 3000 steps of conjugate gradient were performed to gently relax the water molecule around the protein, which was constrained imposing harmonic position restraints of 1000

kJ mol^{-1} nm^{-2}. In the third and in fourth phases, the protocol was repeated without using any constraint. Subsequently the minimizations steps, the systems were equilibrated in four steps. In the first stage, the systems were slowly heated up from 0 K to 300 K in 1 ns at 1 atm, followed by two stages of equilibration in isothermal-isobaric (NPT) conditions for 600 ps, in which positional constraints were applied on the solute atoms (force constant of 1000 kJ mol^{-1} nm^{-2}). In the last stage, a 400 ps unrestrained MD simulation was performed at 300 K and 1 atm to assess the stability of the systems. The temperature and pressure were regulated by a Berendsen thermostat and barostat, respectively (Berendsen et al., 1984). The LINCS algorithm (Hess et al., 1997) was used to constrain all bonds involving hydrogen atoms and the time step was set to 2 fs Periodic boundary conditions (PBC) were applied. The particle mesh Ewald method (PME) (York et al., 1994) with a cut-off length of 1 nm was used to treat long-range electrostatic interactions, and the same cut-off length was adopted for van der Waals interactions. Finally, all monomer systems were simulated for 170 ns while dimer systems were simulated for 200 ns. For the dimer systems, we decided to carry out longer MD simulations (ca. 200 ns) because the equilibrium state of the two dimer systems was reached at ca. 110 ns. To sum up, the AGT-Mi protein was simulated, in all its mutated forms, totalizing ca. 740 ns of simulation time. For the production runs, a Nose-Hoover thermostat (Hoover, 1985 and Nosé, 1984) at 300 K and an Andersen-Parrinello-Rahman barostat (Parrinello and Rahman, 1981 and Nosé and Klein, 1983), at 1 atm were used. The trajectories were analyzed using the standard GROMACS tools. The root mean square deviations (RMSD) and the root mean
square fluctuations (RMSF) were performed through GROMACS inbuilt tools. RMSD plots for the AGT-Mi and F152I-Mi mutant were calculated without considering 4-23 region, which it is extremely flexible in the monomeric form and introduces artifacts to RMSD analysis. Cut-off of 4 Å between N-O atom pairs as a definition of salt bridge formation between Arg118 and Asp243 was used (Donald et al., 2011). All the graphical displays were generated using XMgrace program. The molecular graphics shown in this work were created using UCSF Chimera (Pettersen et al., 2004).

Essential Dynamics (ED). ED was performed for AGT-Mi and F152I-Mi MD simulations according to principal component analysis (PCA). ED is a standard technique able to highlight the dominant motions of the investigated systems (Kumar et al., 2013 and Rajith et al., 2014). The principal components (PCs), also called eigenvectors, were calculated by using ED and represent the protein's collective motions. Here, the first eigenvector, that accounts for ca. 50% of the total motions accomplished by the system, was used to understand the dynamical differences between the AGT-Mi and the F152I-Mi systems. Therefore, RMSF along the first eigenvector for the AGT-Mi and F152I-Mi were calculated. Covariance matrix was calculated using Ca atoms and the matrix was diagonalized. The PCs were extracted from the covariance matrix. PCA was performed in three steps: i) the covariance matrix was calculated using the Ca atoms; ii) the matrix was diagonalized and iii) a set of eigenvectors and eigenvalues were extracted from the covariance matrix. The eigenvectors correspond to a direction in the multidimensional space while the eigenvalues express the amplitude of the motion of each eigenvectors. PCA was carried out by using the standard GROMACS tool i.e. g_covar and g_anaeig.

6.4 Results

Our collaborators guided by bioinformatic information and experimental data identified Arg118 as a key residue playing a relevant role in the dimerization of AGT-Mi (Dindo et al., 2016). Thus, we hypothesized that some PH1-disease causing mutations could be associated to mutated aminoacids near to or interacting with Arg118 and a search in this direction has been made. In the 3D structure of AGT, considering an ideal sphere of 15Å radius whose center is the Cα of Arg118 we noticed the presence of several aminoacids whose mutation whose mutation is associated with PH1. These aminoacids are found in the large domain of the protein. It can be hypothesized that the pathogenicity of these mutations is due to drastic perturbations of the microenvironment of Arg118 and/or Asp243. We concentrated our attention on the mutations F152I and I244T which are among the most common and the most studied in PH1. Therefore, the effect of F152I and I244T pathogenic mutations was investigated by using extensive MD simulations on the modeled structures of AGT-Mi, I244T-Mi and F152I-Mi in the dimeric form.

6.4.1 I244T-Mi mutant

The comparison between AGT-Mi and the I244T-Mi mutant showed that, on the basis of RMSD plots, the simulations reached convergence at ca. 110 ns (Fig. S6.1), thus providing a solid base for further analysis.

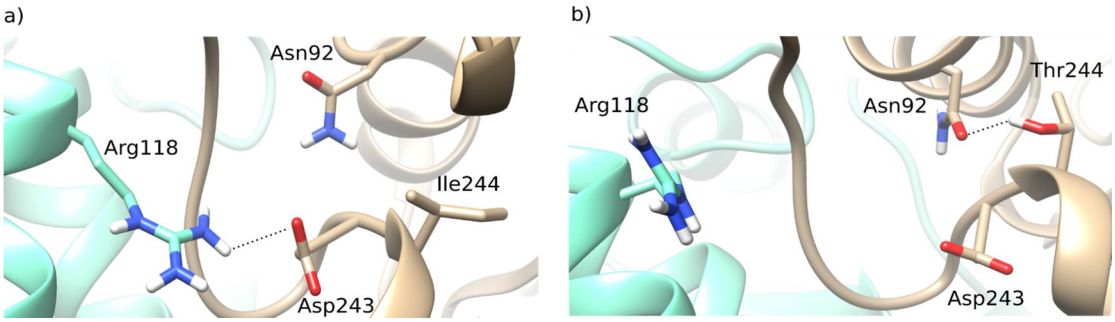


Figure 6.1. Perturbation of the microenvironment of Arg118. Panel **A**: Arg118-Asp243 salt bridge, together without Asn92-Thr244 hydrogen bond interaction. Panel **B**: breaking of Arg118-Asp243 salt bridge, together with Asn92-Thr244 hydrogen bond interaction; The two monomeric subunits are displayed in aquamarine and brownish. Depicted hydrogen atoms were added within the setup of the simulations. Reprinted from Dindo et al., 2016.

By visual inspection, we discover that, over 200 ns of simulation, the electrostatic interaction formed by Arg118 and Asp243 is maintained for ca. 87.3% of the time in AGT-Mi, while it remains stable for just about 17.5% of the time in the I244T-Mi mutant. Thus, we suppose that the I244T mutation could have a straightforward influence on the formation of a key interface contact. Indeed, we observed that the breaking of the Arg118-Asp243 ionic bond corresponds with the formation of an hydrogen bond interaction between Thr244 and Asp92, which provokes a backbone conformational change (Fig. 6.1 Panel A and B). Conversely, in AGT-Mi, this fact is not possible considering that the Ile244 cannot form H-bonds interactions with neighboring residues, assuring the stability of the Arg118-Asp243 ionic interactions.

6.4.2 F152I-Mi mutant

On the other hand, as regards the F152I-Mi mutant, we did not notice any significant changes between the interface contacts of dimeric AGT-Mi and F152I-Mi along the simulations. Thus, we perfomed MD simulations for 170 ns on the monomeric form of both AGT-Mi and F152I-Mi.

For the two monomeric systems we have analyzed i) the root mean square deviations (RMSD) to understand how the systems evolve from the initial structure and ii) the residues' root mean square fluctuations (RMSF) in order to study how flexible they are along the simulations (see section 6.3 for more details). From the RMSD analysis, it can be noticed that both systems became equilibrated after ca. 10 ns to values of about 0.25 nm (Fig. S6.2). RMSF data on both systems showed high fluctuations at the N-terminal and C-terminal loop extremities and in loops that are far from the dimer interface and exposed to the solvent e.g. 252-255, 280-285 and 355-357 (Fig. S6.3). However, neither these analyses nor visual inspection helped us to discover important perturbations on the dimer interface portions between AGT-Mi and F152I-Mi mutant. Thus, we decided to carried out further studies of essential dynamics (ED) on the MD simulations.

ED technique allows both, to highlight the dominant motions of the system, which are often meaningful and correlated with protein function, and reduces the complexity of the data. Fluctuations analysis by ED indicated that the F152I mutation could change the dynamics of several regions of AGT-Mi, i.e., 95-97, 118-121, 141-145, 171-176 and 281-290, with peaks for residues 97, 118, 143, 175 and 281. These results indicate that the latter aminoacids have a different behavior in terms of flexibility in F152I-Mi compared with AGT-Mi. Indeed, it can be noticed (Fig. 6.2) that one of the regions that presents different behavior in the mutated system spans aminoacids 118-121, located at the dimerization interface. As result of this, we could hypothesize that the F152I mutation could affect dimer formation by increasing the flexibility of the region 118-121, thus possibly reducing the complementarity between the two subunits. Moreover, roughly 70 % of all cumulative protein fluctuations are comprised within the first two principal components (PCs), so the overall motion of the proteins in phase space can be showed as a projection of the first two eigenvectors (PC1, PC2). Our results show (Fig. S6.4) that F152I mutant covers a larger region of conformational space than AGT-Mi, suggesting an overall escalation in the flexibility of the mutant form than the AGT-Mi. The values indicating the total variance of the systems for AGT-Mi and F152I-Mi mutant protein were 28.1565 nm^2 and 52.9434 nm^2 respectively, validating the hypothesis of the overall increased flexibility of F152I-Mi mutant than AGT-Mi.

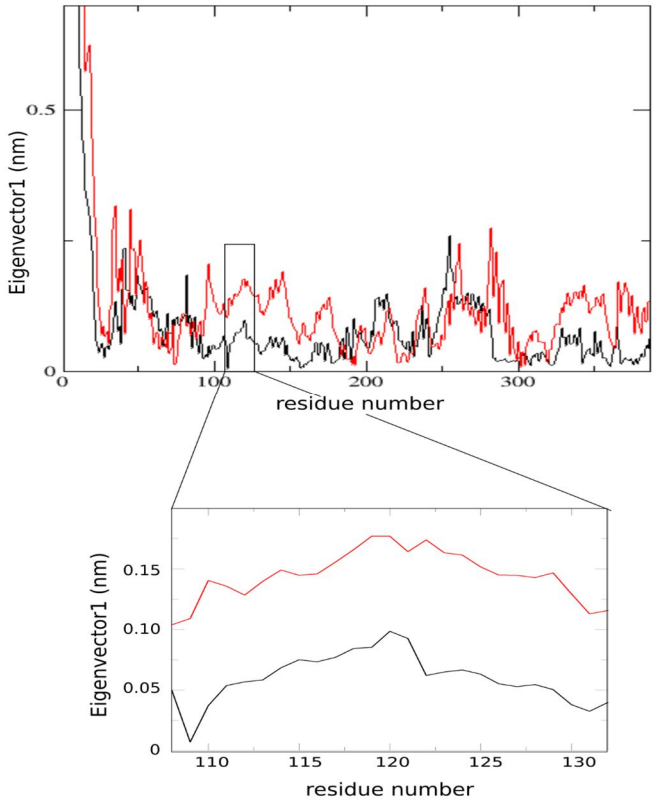


Figure 6.2. RMSF of Cα atoms along first eigenvector. The simulations were run on AGT-Mi (black) and F152I mutant (red) monomer forms. In the box, we focus our attention on the RMSF value of the residues in the dimer interface. Reprinted from Dindo et al., 2016.

6.5 Discussion and Conclusions

Understanding the AGT dimerization process is important for two reasons. In the first place, it allows us to gain a better insight into the catalytic mechanism of the enzyme and, in the second place, it helps clarifing why some PH1-causing variants encoded on the AG-Mi, i.e. F152I and I244T, are mistargeted into mitochondria. The import into peroxisomes mainly operates on a fully-folded dimeric state of AGT (Cellini et al., 2010), while that in mitochondria needs a monomeric state of the protein provided with a mitochondrial targeting sequence resulting from the P11L polymorphism. Since most of glyoxylate is synthesized in the peroxisomes, only peroxisomal AGT is metabolically efficient. Hence, both the dimerization and the correct subcellular localization of AGT are functionally crucial events for this protein. It has been hypothesized that the effect of some aminoacids variations together with P11L could provoke a destabilization of the AGT-Mi dimeric structure (Cellini et al., 2010).
Anyhow, the effect of these mutations from the structural point of view has not been studied yet.
Phe152 and Ile244 aminoacids are not located in the dimer interfacial, anyway, it cannot be excluded that substitution of Phe152 and Ile244 with Ile and Thr, respectively, might indirectly alter/reduce the dimerization mechanism. Thus, in order to verify this theory,

extensive MD simulations were carried out on AGT-Mi, F152I-Mi and I244T-Mi. The computational results highlighted that both mutations hinder the dimerization mechanism. Indeed, replacement of Thr244 with Ile leads to the breakage of the key strong salt bridge between Arg118 and Asp243 and the formation of a new hydrogen bond interaction between Arg118 and Asp92. On the other hand, in the F152I-Mi monomer, but not in the dimer, the region spanning aminoacids 118-121 results to be more flexible than AGT-Mi monomer. Thus, two hypotheses can be advanced: 1) the buttressing of monomer to form a dimer might prevent increased movement at the dimer interface, and 2) the high flexibility of residues 118- 121 could result in an attenuated monomer in its dimerization process. In addition, in the paper, a new model for the dimerization mechanism of AGT-Mi, I244T-Mi and F152I-Mi was advanced taking into account the results of this research and the previous studies.[5] The study was carried out in cooperation with the group led by Prof. Carla Voltattorni.

[5] For the model details please see Dindo et al., 2016.

Chapter 7

Modeling human ileal bile-acid-binding protein/membrane interactions within a crowded cellular environment

This chapter gives the argumentation published in Ceccon et al., 2015.

7.1 Introduction

Operation of the cellular machinery relies on finely tuned dynamic association/dissociation equilibria among the molecules of life (Alberts et.al., 2002). The type of interactions between macromolecule depend on the chemico-physical characteristics of their molecular surfaces (Keskin et.al., 2008). Over the years we have reached a deep knowledge about the factors of formation of long-lived protein complexes, leading to a full description of interface architectures (Nooren and Thornton, 2003; Janin et.al., 2008; Keskin et.al., 2008 and Kastritis and Bonvin, 2013). Anyhow, in recent times, it has been also demonstrated that short-lived interactions are important in regulation of biochemical pathways and signaling cascades in the cells (Ozbabacan et.al., 2011 and Perkins et.al., 2010). Transient multimolecular associations forming dynamic protein assemblies have been recognized as fundamental components of cellular organization and signaling (Nussinov and Jang, 2014). Recent investigations have allowed to explain at atomic-level the ultraweak affinity complexes governing biological energy transduction, enzyme catalysis, and transient self-assembly (Crowley and Ubbink, 2003 and Xing et.al., 2014). Presumably, the quantity of ultraweak interactions within the cell competes with that of stable interactions, however our knowledge of loose macromolecular interactions is still limited.

The cell interior is a complex heterogeneous milieu (Alberts et.al., 2002). Indeed, a macromolecule can interact with other macromolecules within the cell and this has a significant influence in its conformational distributions, binding propensity, diffusional transport (Gierasch and Gershenson, 2009). It should be noted that properties and interaction of macromolecules in biochemical experiments *in vivo* are usually studied in dilute solutions (lower than 0.1 g/L), where the effect of crowding can be neglected. However, to understand the biochemical processes, which proceed in the living cell, the crowding conditions should be simulated in experiments *in vitro*. In principle, the peculiar characteristics of cellular environments are four: 1) macromolecular crowding, 2) local viscosity, 3) compartmentalization and 4) confinement (Zhou et.al., 2008).

The number of intermolecular encounters as well as the probability of shifting of the position of binding equilibria are of course augmented if we think to the high total concentration of macromolecules and the low available space in the cell. Thus, crowding and the other factors actively govern short-lived associations and optimize the competition between functional and casual interactions. In this case, weak nonspecific interactions cooperate to support or hinder functional intracellular communication. In the cell, macromolecules also called crowding agents occupy a significant part of the total volume of the medium, (a total of 400 g/L, which corresponds to a 20-40% occupation of

total cellular volume). Therefore, to simulate macromolecular crowding *in vitro*, crowding agents should be added to the solution. The essential features required for the macromolecular crowders are three: 1) their molecular mass should range from 50 to 200 kD; 2) the molecules should have the shape of globules to avoid making the solution too viscous, 3) the crowding molecules should not have the ability for specific interactions with the molecules under test because crowding implies only steric repulsion between different molecules. The most commonly used crowding agents are BSA, ovalbumin, Ficoll 70, and polyethylene glycol (PEG).
In general, macromolecular crowding effects stem from steric (hard) repulsions and weak chemical (soft) interactions (Minton, 2013) Association reactions and protein stability are entropically favored in crowded solutions due to excluded volume. Soft repulsions usually augment the effects of steric hindrance.
Hard repulsion has been successfully studied by synthetic crowding agent e.g. Ficoll and dextran, which have exhibited a behavior similar to inert cosolutes (Zhou et.al., 2008).
On the other hand, biomolecular crowding agents can significantly interact with test proteins (Wang et.al., 2012). Nonspecific attractive interactions between a protein and the crowder sometimes provoke partial destabilization, aggregation or misfolding. Furthermore, the protein mobility can be altered by transient interactions with cytoplasmic components (Kuznetsova et.al., 2014 and Sanfelice et.al., 2012). This hypothesis was confirmed by in-cell (NMR) spectroscopy researches, which showed that the signals of some globular proteins remained hidden as the result of formation of transient high molecular weight complexes (Sanfelice et.al., 2012 and van den Berg et al., 1999). Diffusional transport of cytosolic proteins strongly depend on the environment. The latter is bounded by a diversity of surface boundaries comprising the plasma membrane (accounting to about 700 μm^2 in an average cell), internal membranes (7000 μm^2), and the cytoskeleton (94000 μm^2) (Wang et.al., 2010). Consequently, the cytosolic proteins can probably interact with a high number of membrane lipids, in addition to cytoskeletal proteins. The largest class of lipids in most eukaryotic membranes is Phosphoglycerides, in which phosphatidylcholine (PC) is more than 50% of the phospholipids (Wang et.al., 2011). PC molecules have generally a saturated and a *cis*-unsaturated C_{16}/C_{18} fatty acyl chain. The type of protein-membrane interactions is due to the chemical and physical characteristics of the lipid components and of the aminoacid residues. In the case of peripheral (or non-permanent, amphitropic) membrane proteins the transient anchoring to the lipid bilayer is mostly influenced by the identity of the lipid head groups rather than by the structure of the hydrophobic core (Crowley et.al., 2011). The bulk lipids in the cytoplasmic leaflet of cellular membranes are electrically neutral (e.g. cholesterol), have zero net charge (e.g. PC, in which the choline head group is zwitterionic) or are anionic (e.g. phosphatidylserine, PS) (Wang et.al., 2012). It can be anticipated that structure with many basic residues on their interface experience electrostatic interactions on approaching the lipid layer.
In this chapter, we study the propensity for weak chemical interactions of a cytosolic protein (human ileal bile acid binding protein), used as test, with membrane cellular components (POPG) by using computational approaches.

Human Ileal Lipid-Binding Protein. Bile acids are generated from cholesterol in the liver and they are secreted with bile into the small intestine, where they aid in the digestion and resorption of fat and fat-soluble vitamins (Hofmann, 2011). Most bile acids (95%) are reabsorbed at the end of the small intestine in the ileum. After absorption, they are

recirculated back to the liver with portal blood. These recirculated bile acids are able to inhibit a cholesterol hydrolase, which is the rate limiting enzyme within the conversion cascade from cholesterol to bile acids. Hence, this bile acid recycling process is an important regulator for serum cholesterol levels and thus of special medical and pharmacological interest. Using photoaffinity-labeling experiments, it was shown that an integral membrane protein of 93 kDa and a peripheral protein of 14 kDa are involved in bile acid transportation along the duodenum-ileum axis (Kramer et.al., 1993). The 15-kDa protein has been identified as ileal lipid-binding protein (ILBP, also known as Gastrotropin or ileal bile acid-binding protein (IBABP)). ILBP is a cytosolic lipid binding protein that binds both bile acids and fatty acids. The secondary structure of free ILBP (Fig 7.1) consists of two short α helices: helix I: Asn13-Leu21, helix II: Ser 25-Arg32, and 10 βstrands, and cover the residues Gly4-Met8 (strand A), Ile36-Gln42 (strand B), Phe47-His52 (strand C), Thr58-Phe63 (strand D), Ser69-Gln72 (strand E), Thr78- Met85 (strand F), Leu90-Asn93 (strand G), His98-Ile103 (strand H), Leu108- Thr113 (strand I), and Thr118-Arg125 (strand J). The adjacent βstrands are linked either by β turns (strands B-C, F-G, G-H, H-I, I-J) or by a turn of five residues. The tertiary structure of human ILBP is similar to other lipid-binding proteins, whereas there are significant differences on ligand binding within the inner core. The 10 antiparallel β strands of human ILBP are arranged in two nearly orthogonal β sheets forming a so called βclam structure (Sacchettini et al., 1988) with a gap between βstrands D and E. The two βhelices close this βbarrel on one side. The global fold of human ILBP is typical for all lipid binding proteins including the large family of fatty acid binding proteins (FABP). This structural motif is similar in the free and complexed ILBP form and was classified as 10 stranded meander β sheet topology folded on itself in the SCOP database (Murzin et al., 1995).

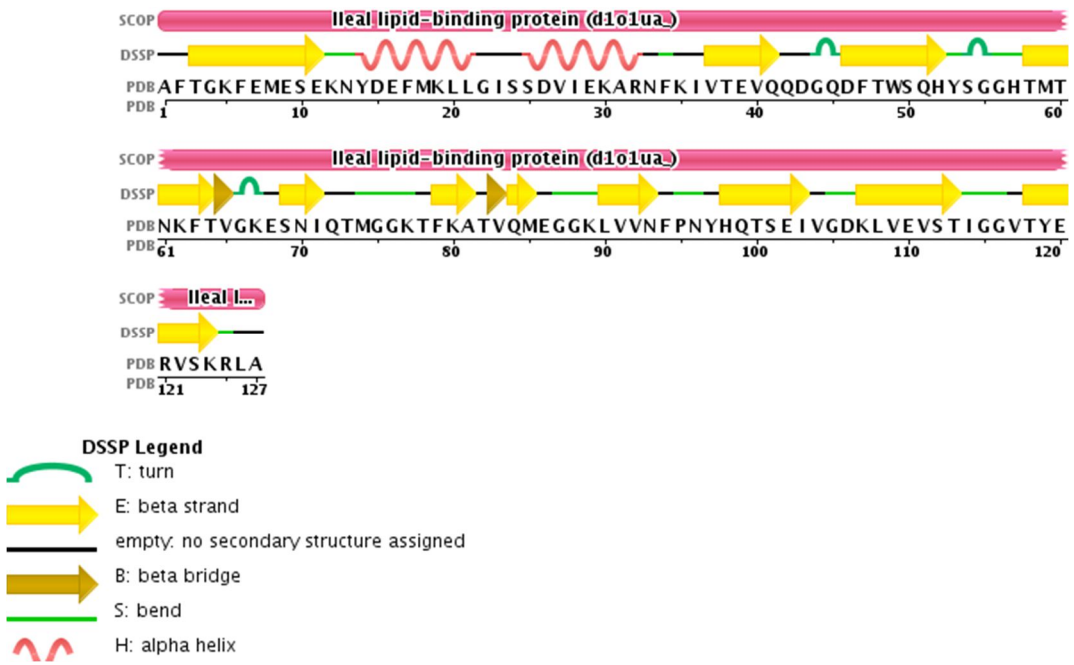


Figure 7.1: Sequence Display for the Entities in PDB 1O1U.

IBABP belongs to the family of intracellular lipid binding proteins (iLBPs). The latter is important for the intracellular lipid transport and the systemic metabolic homeostasis.

Furthermore, they are largely distributed among tissues and they have a high intracellular concentration (ca. 7-11% of total cytosolic proteins) (Schaap et al., 2002 and Vergani et.al., 1990)[6].

7.2 Aim of the research described in this Chapter

- To gain structural insights into the dynamics of IBABP protein and POPG membrane interactions by MD simulations.

7.3 Material and Methods

Coarse-grained MD setup. All MD simulations described here were carried out using the software GROMACS 4.5.4 (Hess et.al., 2008) and the MARTINI coarse-grained (CG) force field version 2.0 (Marrink et.al., 2007). The ElNeDyn model (Periole et.al., 2009) was used to construct the CG models of the protein. The initial coordinates of IBABP (PDB id: 1O1U) (Kurz et al., 2003) were converted to CG representation, using the *martinize.py* script available on the MARTINI force field web site. The secondary structure for the protein was obtained by using the DDSP program (Touw et.al., 2015 and Kabsch and Sander, 1983). The CG protein was then solvated in standard MARTINI water molecules. The system was minimized by 1,000 steps of steepest descent optimization and then two 4 ns-long steps of solvent equilibration in the NPT ensemble with 10 fs and 20 fs time steps, respectively. In the equilibration runs the temperature was held at 300 K by using separate velocity rescale thermostat (Bussi et al., 2007) and the pressure was maintained at 1 bar using the Andersen-Parrinello-Rahman barostat (Parrinello and Rahman, 1981). Periodic boundary conditions were applied and a shifted potential was used to calculate electrostatic interactions. The cut-off values for the real part of the electrostatic interactions and for the van der Waals interactions were set to 1.2 nm. The insane.py script available on the MARTINI force field web site (http://md.chem.rug.nl/) was used to create a fully hydrated bilayer of POPG with physiological ion concentration (0.150 M). The area per lipid was set up to 53 $Å^2$ and the final membrane consisted of about 392 POPG molecules in an area of 11 × 11 nm. The system was minimized and equilibrated in 4 ns to relax the initial grid order imposed by insane.py. The pre-relaxed protein and membrane were centered at 5 × 5 × 7 nm and at 5 × 5 × 15 nm, respectively, in a virtual box of ca. 11 × 11 × 18 nm. In this way the minimum distance between IBABP and the membrane was ca. 3.5 nm. Therefore, the final system was solvated using standard MARTINI water molecules with 11,414 CG water beads (equivalent to 45,656 water molecules). The water molecules overlapping with the lipid bilayer were removed. The obtained system was used as input to generate two systems with different ionic concentrations. In the system at higher ionic concentration (0.175 M), 617 Na^+ and 224 Cl^- ions were placed by replacing randomly chosen water beads

[6] In this chapter, the results and the experimental analysis carried out on the other cytosolic proteins (bovine serum albumin, ubiquitin and hen egg-white lysozyme) are not discussed, so for further information please see Ceccon et al., 2015 paper.

throughout the simulation box to obtain a neutralized system. For the system at lower ionic concentration (0.030 M), 431 Na^+ and 38 Cl^- ions were used.

Details of the CG MD simulations. Each system was energy minimized by performing 1000 steps of the steepest-descent algorithm. The two systems were then equilibrated for 4 ns by using a LeapFrog integrator with a time step of 20 fs in the NPT ensemble. The temperature was kept at 300 K by weak coupling to an external heat-bath using the velocity rescale thermostat with a with a relaxation time of 1 ps (Bussi et. al., 2007). Lipids, protein and water were coupled separately to avoid heat flow. The pressure was fixed at 1 bar with a coupling constant of $\tau_p = 1.0$ ps and compressibility of 3.0×10^{-4} bar^{-1} using the Berendsen barostat (Berendsen et al., 1984). Periodic boundary conditions were applied. Electrostatic interactions were calculated using a shifted potential with a cut off of 1.2 nm and a dielectric constant of 15. Van der Waals interactions were also calculated using a shifted potential with a cut off of 1.2 nm and a switch at 0.9 nm. For the production runs, the pressure was maintained constant by weak coupling to an external reference pressure of 1.0 bar using a Andersen-Parrinello-Rahman barostat (Parrinello and Rahman, 1981) with a time constant of 12 ps. For each system at different ionic concentration, twenty independent 4 μs-long MD simulations differing only for the initial velocities were performed, for a total of 80 μs. The trajectories were sampled every 1000 ps and analyzed using the standard GROMACS tools. In particular, the distance between protein and membrane and the residues mostly interacting with the lipids throughout the simulations were followed by using *g_mindist* program. IBABP was considered in contact with the lipid bilayer if the protein-lipid distance is less or equal to 0.7 nm. This distance was chosen taking into account that MARTINI's beads have an average effective radius of 0.47 nm (Marrink et al., 2007 and Marrink et al., 2004). The d_{Trp49} distance was sampled every 500 frames.

7.4 Results

Molecular simulations can provide a powerful and complementary insight into the dynamics of IBABP protein and POPG membrane interactions. On the other hand, biomolecular dynamics occur over a wide range of time scales, and the choice of approach to study them depends on the peculiarity of the studied system (Lindahl, 2008). Considering that the diffusion coefficient of the protein (ca. 30 kDa) is ca. 100 $\mu m^2/s$ (Eliseo et.al., 2007), and a distance of IBABP from the membrane of ca. 3.5 nm (see below), then the interaction between IBABP and POPG can be estimated to occur on the boundary between the ns and the μs time scale. Such time scale is accessible to atomistic MD simulations, but it is computationally expensive. (The time scale (τ) for a particle to travel a distance x is given on the average by $\tau \approx x^2/D$, where D is the diffusion coefficient for the particle.) Thus, here we take advantage of the less expensive CG MD and of the MARTINI force field (Marrink et.al., 2007) to perform twenty 4 μs-long MD simulations of the IBABP-POPG interaction considering two different ionic concentrations mimicking the highest and the lowest concentrations used in the NMR experiments (namely we set NaCl concentration to 0.030 and 0.175 M, respectively). In total, we performed 160 μs of CG MD. In our computational setup, we positioned IBABP center of mass at 8 nm from the center of mass of an equilibrated POPG bilayer (Figure 7.2 A), thus putting the protein and the membrane at a minimum distance of ca. 3.5 nm. This

distance was selected in order to achieve a reasonable compromise between having a starting position of IBABP isolated from the membrane and possibility to observe a contact between IBABP and the POPG bilayer in the simulations time scale. All the simulations started from the same initial geometry but differ for the initial velocities. In order to monitor the distance between IBABP and the membrane along the trajectories, we measured the minimum distance between Trp49 and POPG (d_{Trp49}). In our simulations, d_{Trp49} drops to values below 2 nm between 10 and 685 ns (average time ca. 175 ns; Figure 7.2 B,C and Fig.S7.1, Fig.S7.2), indicating a contact time of IBABP with the membrane in pretty agreement with our estimations. Before contacting the membrane, IBABP remains in a conformation similar to that observed in the original structure (backbone RMSD between 0.1 and 0.2 nm). Before the contact with the membrane, IBABP structure does not undergo significant structural rearrangements (average RMSD = 0.13 nm with respect to the starting structure). After the contact with the membrane, the IBABP structure can adopt a slightly different conformation, as evidenced by the distribution of the RMSD values in function of the IBABP minimum distance from the membrane (Fig.S7.3). In particular, the protein in contact with the POPG bilayer can adopt an additional conformation (RMSD ca. 0.17 nm) that is only poorly populated in solution. This conformation is characterized by a different orientation of the loop joining strands βC and βD and helix αII. Minor fluctuations can be also observed in the loop joining strands βG and βH, as well as in strands βF and βJ (Fig.S7.4). d_{Trp49} and visual inspection of the trajectories were used to distinguish between different binding poses of IBABP on the POPG bilayer. In the most (65%) of the simulations at the lowest ionic concentration, the protein interacts with the membrane using one surface region (called patch 1 hereafter, see Table 7.1) identified by the shortest average distance between Trp49 and the membrane ($<d_{Trp49}>$ ca. 0.8 nm) and by a large surface of interaction (28% of the total protein surface). Patch 1 is located next to strands from βC to βF (Figure 7.2 D) and is characterized by a slightly positively charged surface. At the lowest ionic concentration, IBABP appears to interact with the membrane also using two less populated interaction surfaces: patch 2 and 3 in Table 7.1, observed in the 25% and 10% of the trajectories, respectively (see Table 7.1, Figure 7.2 D). Both patches 2 and 3 are negatively charged. The three interaction patches observed at the lowest ionic strength are observed also in the simulations at the highest ionic concentration. On the other hand, IBABP contacts the membrane using patch 1 only in the 50% of the simulations, patch 2 in the 20% of the cases and patch 3 together with two new binding poses (patch 4 and 5 in Table 7.1) are observed in the 10% of simulations each. Patch 3 and 4 are characterized by a negatively charged surface. Interestingly, at the highest ionic concentration the binding pose can interconvert along the trajectory (see Figure 7.2 C bottom panel and SI); an event not observed at the lowest ionic concentration, where IBABP retains the same orientation on the membrane for the entire duration of the simulations. In conclusion, our CG MD simulations showed that, if IBABP is placed in the proximity of the membrane (i.e.: d_{Trp49} less than 10 nm) then it always comes in close contact with the membrane in a time between ca. 10 and 900 ns. In our model at low ionic strength, IBABP interacts with the membrane using preferably a large positively charged region on the surface of the protein and such interaction is stable in the μs time scale. On the other hand, at higher ionic strength, the IBABP interaction with the membrane can occur by using a larger number of surface patches that can interconvert in the simulations time scale. Interestingly, the IBABP patch most frequently used to contact the POPG bilayer at both ionic concentrations is positively charged, while all the other observed patches are

negatively charged. This indicate that the IBABP/POPG interaction is partially driven by electrostatics (the POPG surface is negatively charged), but that also other factors that are difficult to characterize at the atomistic level due to the nature of the model used can contribute to the formation of the protein/membrane encounter complex. The interaction was to some extent specific at low ionic strength and unspecific at high salt concentration. CG MD simulations support the finding (based on fluorescence spectroscopy) that the major docking site was formed by residues in the sheet domain rather than the helix–turn–helix motif (Ceccon et.al., 2015).

Table 7.1. IBABP surface regions (patches) found in contact with the membrane during the CG MD simulations.

Patch#	Surface residues	$\langle d_{Trp49} \rangle$ (nm)	IBABP surface (%)
Patch1	Ser24, Asp26, Val27, Asp43, Thr48, His52, Tyr53, Ser54, Gly55, Gly56, His57, Thr58, Met59, Thr60, Lys62, Thr64, Lys67, Glu68, Ser69, Asn70, Ile71, Gln72, Gly76, Lys77, Thr78, Phe79, and Lys80	0.8	28
Patch2	Glu9, Ser10, Glu11, Lys12, Asn13, Asp15, Glu16, Arg32, Asn33, Asp106, Lys107, Glu120, Val122, Lys124, and Leu126	1.5	18
Patch3	Ala1, Phe2, Thr3, Gly4, Lys5, Gln41, Gln42, Asp43, Gly44, Gln45, Gly87, Gly88, Gly105, Asp106, Arg125, and Leu126	1.0	18
Patch4	Glu16, Lys19, Leu20, Leu21, Gly22, Ile23, Ser24, Ser25, Glu68, Asn70, Gly75, Gly76, Lys77, Thr78, Phe79, Lys80, Ala81, Thr82, Asn93, Phe94, Pro95, Asn96, Tyr97, His98, Ile114, Gly115, Gly116, and Val117	1.7	26
Patch5	Ala1, Phe2, Thr3, Gly4, Lys5, Phe6, Glu9, Ser10, Lys35, Ile103, Val104, Gly105, Asp106, Lys107, Val122, Lys124, Arg125, Leu126, and Ala127	1.5	19

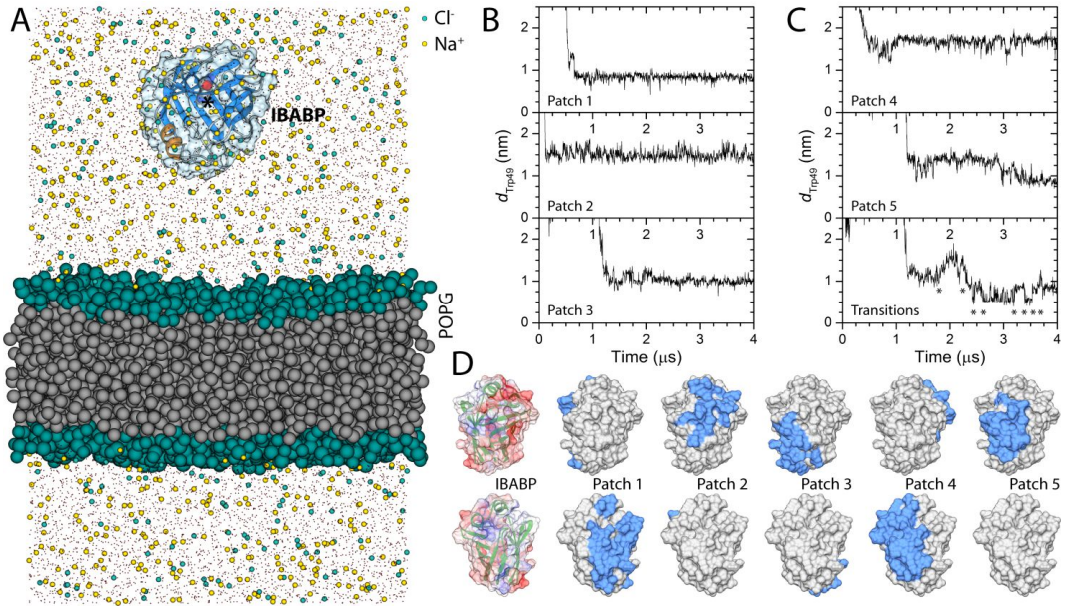


Figure 7.2. (A) Simulation box used in the CG MD simulations at higher ionic concentration. IBABP is reported by using a ribbons diagram colored according to the secondary structure and the transparent solvent excluded surface. Trp49 is shown as balls model and is highlighted by a * symbol. The membrane is depicted as grey and dark green spheres for the hydrophobic and polar CG beads, respectively. The CG water molecules are indicated by red dots while Na^+ and Cl^- ions are reported as yellow and cyan spheres. (B,C) Time evolution of d_{Trp49} along selected simulations at the lowest (B) and at the highest (C) ionic concentrations. Transitions between different patches are highlighted by * symbols. (D) Molecular surface of IBABP colored according to the surface electrostatic potential contoured from -5.0 (intense red) to +5.0 kT/e (intense blue, where k is the Boltzmann constant, T is the absolute temperature, and e is the electron charge) (left panels) and to the residues found in contact with the membrane during the CG MD simulations. The surfaces in the bottom panels are rotated by 180 degrees along the vertical axis with respect to the surfaces depicted in the top panels. Reprinted from Ceccon et.al., 2015.

7.5 Discussion and Conclusions

In this chapter, we focused on transient protein-membrane interactions, as might be occurred in a crowded cytosolic milieu. Human IBABP was used as test protein. The latter is present in the cell in high concentration and so far, their recognition surface patches have not been studied. We exploited large unilamellar vesicles to mimic lipid membranes. The overall slightly negative IBABP macromolecules (pI = 6.5) adsorbed weakly to anionic vesicles. The electrostatic nature of the IBABP-membrana interaction was described by computational experiments, which were performed by using CG MD simulations. This was then confirmed by repeating experimental binding assays at different ionic strength or in the presence of neutral vesicles. In addition, MD simulations enable us to understand the surface patches putatively involved in the collisions with the membrane. The poorly defined protein molecular orientations at the lipid surface supported the view of unspecific IBABP-membrane binding. The ultraweak interactions of IBABP with membrane mimics exemplify potentially interfering associations within functional intracellular communication pathways and point out subtle effects of cosolutes

on protein conformational dynamics. The study was devised in association with NMR group lead to Prof. Michael Assfalg.

Chapter 8

Conclusions

In the last years, with continuing advances in the methodology and speed of computers, which is still doubling every eighteen months or less (according to Moore's law), MD studies are being extended to larger system, greater conformational changes and longer time scales. This prompted the scientific community to consider MD simulations of biological molecules as virtual experiments, which make possible the investigation of motions that have particular functional implications and to obtain information that would not accessible from wet-lab experiments. A great challenge still to be solved is to observe functionally important biological events that typically occur on time scales of many microseconds or milliseconds. One way of overcoming this problem is the use of CMB approaches. Another important challenge is represented by the lack of structural information on many proteins of outstanding biological importance. This challenge can be undertaken by combining state of the arte *in silico* techniques i.e. homology modeling, MD simulations and docking, with experimental protocols. Here, I have undertaken both challenges with the aim of characterizing, in collaboration with experimental groups, different biological systems.

Our initial studies comprise the computational design and optimization of MIPs. The first step of my project was to design new protocols for the predictive rational design of MIPs for practical applications. In recent years, the Polymer Group led by Professor Piletsky at Leicester, UK has used molecular modeling (using Sybyl (Tripos Inc., St. Louis, MO, USA)) in the optimization of polymer compositions to make high affinity synthetic receptors based on Molecular Imprinting.

We investigated state of the art programs such as HADDOCK (Dominguez et. al., 2003 and van Zundert et. al., 2016), Autodock (Morris et. al., 2009), Vina (Trott and Olson, 2010) and GROMACS (Pronk et al.,2013) (and their parameters were optimized for MIP design) as novel and alternative tools to Sybyl package (Tripos Inc., St. Louis, MO, USA). This study stems from the desire to find a free alternative computational method to the Leicester protocol that relies on purchasing Sybyl (approximately £4,000 per annum).

The first part of the HADDOCK (Dominguez et. al., 2003 and van Zundert et. al., 2016) study was the molecular modeling of MIPs for N terminal of the hepcidin-25 human peptide as a template. Our new HADDOCK (Dominguez et. al., 2003 and van Zundert et. al., 2016) based approach allowed the introduction of constraints arising from the knowledge of the chemical-physical properties during the docking procedure setting some parameters. Our HADDOCK (Dominguez et. al., 2003 and van Zundert et. al., 2016) based protocol was advanced in order to find an alternative method to the LeapFrog algorithm (Tripos Inc., Tripos, St. Louis, MO, USA) to test each individual monomer against the same template. When our HADDOCK (Dominguez et. al., 2003 and van Zundert et. al., 2016) based approach was run against LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA) using the same template (troponin) similar results were obtained. By using our new method, high affinity MIPs for troponin and hepcidin were synthesized (Cenci et al., 2015 and Cenci et al., 2016). Autodock (Morris et. al., 2009) and Vina (Trott and Olson, 2010) were also evaluated for scoring the affinity, location and directionality of the primary bond between a template and FM molecule in a typical apolar organic

porogenic solvent environment. Here, the Autodock (Morris et. al., 2009) scoring function was modified and optimized for the MIPs design. In particular, when our Autodock (Morris et. al., 2009) based approach was run against LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA) using the same template (such as Theophylline and Caffeine) similar results were obtained. Moreover, the computational predictions were also screened against some experimentally determined values, showing good semblance. Thus, an effective short list of FMs candidates can be created with respect to affinity towards the template molecule as well as primary bond location which can be extremely useful when designing imprinted receptor sites with multiple FM. In the same spirit, our collaborators has presented a new Vina (Trott and Olson, 2010) based approach for screening of "virtually imprinted receptors" for rebinding of the molecular template as well as secondary structures, correlating the virtual predictions with experimentally acquired data in three case studies (Bates et al., 2016). Finally, we have also proposed a new GROMACS (Pronk et al.,2013) based protocol to calculate the ideal stoichiometric ratio between the template and monomer (Bates et al., 2017).

As a second step, we have developed a publicly accessible web server named "MIRATE web server" (based on our novel protocols), aimed at the intelligent design of MIPs (http://mirate.di.univr.it/).

In the last stage of this study we wrote a new algorithm capable of simulating the necessary radical polymerization, including chemical and spatial discrimination, hybridization expression, and molecular mechanical minimization for atomic geometry optimization. The novel algorithm, written by using the SPL, was used to effectively identify affinity trends and binding site selectivities were commonly used alternative methods cannot (Cowen et al., 2016). The novel MIRATE web server and the innovative Sybyl script can be used as powerful tools in the design and analysis of MIPs. The MIPs project was possible by a strong collaboration between University of Verona, Leicester and Coruña.

In my second study, we took advantage of state of the art structural bioinformatics tools in order to characterize the HIV-1 Nef/ACOT8 interaction. The latter, indeed, may be important in the endocytosis regulation of membrane proteins and might modulate lipid composition in membrane rafts. While the Nef regions participating to ACOT8 interaction have been already characterize, the structure details of the ACOT8 protein have not been investigated yet. Thus, the lack of structural information hampers the understanding of the functional consequences of the complex formation during HIV-1 infection. In this reasearch, I modelled the ACOT8 and identified the ACOT8 putative contact points involved in the interaction with Nef. Our structural predictions were validated by *in vitro* assays through the development of ACOT8 deletion mutants and enabled us to hypothesize that Nef associates with ACOT8 to avoid the proteasome degradation. These findings improve the comprehension of the association between HIV-1 Nef and ACOT8, helping elucidating the biological effect of their interaction (Serena et al., 2016).

In this thesis, we have also demonstrated the feasibility of computational biophysics instruments (MD techniques), which can be used for different targets. MD simulations is a powerful computational method for delineating motions of proteins at an atomic-scale via theoretical and empirical principles in physical chemistry. This tecnique was used to gain insights into the putative effects of the F152I-Mi and I244T-Mi variants associated with PH1 disease. Our predictions showed that actually both mutations interfere with the dimerization process. In particular, the I244T mutation causes the breakage of the

keystone interaction between Arg118 and Asp243. Conversely, the F152I mutation provokes an increased mobility of one interface region spanning amino acids 118–121, this could attenuate the dimerization process. Our hypothesis paved the way for the design of the AGT folding pathway (Dindo et al., 2016).

Last but not least, we exploited the less demanding CG MD and the MARTINI force field (Marrink et.al., 2007) in order to gain structural insight into the poses of IBABP binding to anionic lipid membranes. In particular, two different ionic strengths were used to reproduce the high and low salt concentrations used in the NMR experiments. Extensive MD simulation showed that IBABP interacts with the anionic lipid membrane in a time interval of 10-900 ns. We identified several IBABP interaction patches and the most frequent contact surface was positively charged. We also detected other less frequent negatively charged patches. This prompted us to hypothesize that the IBABP/POPG attraction was partially driven by electrostatics, but additional forces could contribute to the formation of transient encounter complexes. The interaction was to some extent specific a low ionic strength and unspecific in high salt concentration. The MD simulations outcomes were in agreement with the experimental data based on fluorescence spectroscopy. Indeed, the major docking site included aminoacids in the sheet domain as opposed to the helix-turn-helix motif (Ceccon et al., 2015).

Papers published during my PhD period:

- Ceccon, A., **Busato, M**., Pérez Santero, S., D'Onofrio, M., Musiani, F., Giorgetti, A. and Assfalg, M., 2015. Transient Interactions of a Cytosolic Protein with Macromolecular and Vesicular Cosolutes: Unspecific and Specific Effects. ChemBioChem, 16(18), pp.2633-2645.
- Cenci, L., Andreetto, E., Vestri, A., Bovi, M., Barozzi, M., Iacob, E., **Busato, M**., Castagna, A., Girelli, D. and Bossi, A.M., 2015. Surface plasmon resonance based on molecularly imprinted nanoparticles for the picomolar detection of the iron regulating hormone Hepcidin-25. Journal of nanobiotechnology, 13(1), p.1.
- Cenci, L., Anesi, A., **Busato, M**., Guella, G. and Bossi, A.M., 2016. Molecularly imprinted polymers coupled to matrix assisted laser desorption ionization mass spectrometry for femtomoles detection of cardiac troponin I peptides. *Journal of Molecular Recognition*, *29*(1), pp.41-50.
- Serena, M., Giorgetti, A., **Busato, M**., Gasparini, F., Diani, E., Romanelli, M.G. and Zipeto, D., 2016. Molecular characterization of HIV-1 Nef and ACOT8 interaction: insights from in silico structural predictions and in vitro functional assays. *Scientific reports*, *6*, 22319.
- **Busato, M**. and Giorgetti, A., 2016. Structural modeling of G-protein coupled receptors: An overview on automatic web-servers. The international journal of biochemistry & cell biology, 77, pp. 264-274.
- Dindo, M., Montioli, R., **Busato, M**., Giorgetti, A., Cellini, B. and Voltattorni, C.B., 2016. Effects of interface mutations on the dimerization of alanine glyoxylate aminotransferase and implications in the mistargeting of the pathogenic variants F152I and I244T. Biochimie, 131, pp.137-148.
- Cowen, T., **Busato, M.**, Karim, K. and Piletsky, S.A., 2016. In Silico Synthesis of Synthetic Receptors: A Polymerization Algorithm. *Macromolecular Rapid Communications*, 37(24), pp. 2011-2016.
- Bates F., **Busato M**., Piletska E., Whitcombe M. J., Karim K., Guerreiro A. , del Valle M., Giorgetti A. and Piletsky S. 2016. Computational design of molecularly imprinted polymer for direct detection of melamine in milk. Separation Science and Technology journal. (paper in press)
- **Busato M.***, Distefano R.*, Bates F., Karim K, Bossi A.M., Lopez-Vilariño J.M., Piletsky S., Bombieri N. and Giorgetti A. 2017. MIRATE: MIps RATional dEsign webserver. (paper submitted)
- Bates F., Cela-Pérez M.C., **Busato M.**, Lopez-Vilariño J.M. 2017. Autodock as a virtual screening tool of functional monomers for molecularly imprinted polymers. (paper in preparation)
- **Busato, M.***, Suku, E.*, Mazzocca M., Segato, E., Damre, M., Vargas S.M. and Giorgetti A., 2016. MUnCH-GPCR: ManUally Curated HMM profiles of structurally characterized GPCRs. (paper submitted)

* co-first authors

References:

Alexander, C., Andersson, H.S., Andersson, L.I., Ansell, R.J., Kirsch, N., Nicholls, I.A., O'Mahony, J. and Whitcombe, M.J., 2006. Molecular imprinting science and technology: a survey of the literature for the years up to and including 2003. *Journal of molecular recognition*, *19*(2), pp.106-180.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), pp.3389-3402.

Alvarez, E., Eswar, N., Eramian, D., Webb, B., Shen, M.Y. and Sali, A., 2008. Protein structure modeling with MODELLER. *Structural proteomics: high-throughput methods*, pp.145-159.

Alvarez, E., Girones, N. and Davis, R.J., 1990. Inhibition of the receptor-mediated endocytosis of diferric transferrin is associated with the covalent modification of the transferrin receptor with palmitic acid. *Journal of Biological Chemistry*, *265*(27), pp.16644-16655.

Anderson, G.J. and McLaren, G.D., 2012. *Iron physiology and pathophysiology in humans* (Vol. 704). New York: Humana Press.

Andersson, H.S., Karlsson, J.G., Piletsky, S.A., Koch-Schmidt, A.C., Mosbach, K. and Nicholls, I.A., 1999. Study of the nature of recognition in molecularly imprinted polymers, II: influence of monomer–template ratio and sample load on retention and selectivity. *Journal of Chromatography A*, *848*(1), pp.39-49.

Andrews, P.R., Craik, D.J. and Martin, J.L., 1984. Functional group contributions to drug-receptor interactions. *Journal of medicinal chemistry*, *27*(12), pp.1648-1657.

Ansell, R.J. and Kuah, K.L., 2005. Imprinted polymers for chiral resolution of (±)-ephedrine: understanding the pre-polymerisation equilibrium and the action of different mobile phase modifiers. *Analyst*, *130*(2), pp.179-187.

Arnold, K., Bordoli, L., Kopp, J. and Schwede, T., 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics, 22(2), pp.195-201.

Arold, S., Hoh, F., Domergue, S., Birck, C., Delsuc, M.A., Jullien, M. and Dumas, C., 2000. Characterization and molecular basis of the oligomeric structure of HIV-1 nef protein. *Protein Science*, *9*(06), pp.1137-1148.

Arshady, R. and Mosbach, K., 1981. Synthesis of substrate-selective polymers by host-guest polymerization. *Die Makromolekulare Chemie*, *182*(2), pp.687-692.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Molecular Biology of the Cell, 4th ed., Garland Science, New York, 2002.

Baggiani, C., Anfossi, L., Baravalle, P., Giovannoli, C. and Tozzi, C., 2005. Selectivity features of molecularly imprinted polymers recognising the carbamate group. *Analytica chimica acta*, *531*(2), pp.199-207.

Baker, N.A., Sept, D., Joseph, S., Holst, M.J. and McCammon, J.A., 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, *98*(18), pp.10037-10041.

Bates, P.A., Kelley, L.A., MacCallum, R.M. and Sternberg, M.J., 2001. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins: Structure, Function, and Bioinformatics, 45(S5), pp.39-46.

Beauchamp, K.A., Bowman, G.R., Lane, T.J., Maibaum, L., Haque, I.S. and Pande, V.S., 2011. MSMBuilder2: modeling conformational dynamics on the picosecond to millisecond scale. *Journal of chemical theory and computation*, *7*(10), pp.3412-3419.

Benichou, S., Bomsel, M., Bodeus, M., Durand, H., Doute, M., Letourneur, F., Camonis, J. and Benarous, R., 1994. Physical interaction of the HIV-1 Nef protein with beta-COP, a component of non-clathrin-coated vesicles essential for membrane traffic. *Journal of Biological Chemistry*, *269*(48), pp.30073-30076.

Benkert, P., Tosatto, S.C. and Schwede, T., 2009. Global and local model quality estimation at CASP8 using the scoring functions QMEAN and QMEANclust. Proteins: Structure, Function, and Bioinformatics, 77(S9), pp.173-180.

Bentham, M., Mazaleyrat, S. and Harris, M., 2006. Role of myristoylation and N-terminal basic residues in membrane association of the human immunodeficiency virus type 1 Nef protein. *Journal of general virology*, *87*(3), pp.563-571.

Berendsen, H.J., Postma, J.P., van Gunsteren, W.F. and Hermans, J., 1981. Interaction models for water in relation to protein hydration. In *Intermolecular forces* (pp. 331-342). Springer Netherlands.

Berendsen, H.J., Postma, J.V., van Gunsteren, W.F., DiNola, A.R.H.J. and Haak, J.R., 1984. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, *81*(8), pp.3684-3690.

Berendsen, H.J.C., Grigera, J.R. and Straatsma, T.P., 1987. The missing term in effective pair potentials. *Journal of Physical Chemistry*, *91*(24), pp.6269-6271.

Berweger, C.D., van Gunsteren, W.F. and Müller-Plathe, F., 1995. Force field parametrization by weak coupling. Re-engineering SPC water. *Chemical physics letters*, *232*(5), pp.429-436.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T.G.,Bertoni, M., Bordoli, L. and Schwede, T., 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic acids research, p.gku340.

Brazas, M.D., Yim, D., Yeung, W. and Ouellette, B.F., 2012. A decade of web server updates at the bioinformatics links directory: 2003–2012. Nucleic acids research, p.gks632.

Brooks, B.R., Brooks, C.L., MacKerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G.,Bartels, C., Boresch, S. and Caflisch, A., 2009. CHARMM: the biomolecular simulation program. *Journal of computational chemistry*, *30*(10), pp.1545-1614.

Bruccoleri, R.E., 1993. Application of systematic conformational search to protein modeling. Molecular simulation, 10(2-6), pp.151-174.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S. and Read, R.J., 1998. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallographica Section D: Biological Crystallography*, *54*(5), pp.905-921.

Busato, M. and Giorgetti, A., 2016. Structural modeling of G-protein coupled receptors: An overview on automatic web-servers. The international journal of biochemistry & cell biology.

Bussi, G., Donadio, D. and Parrinello, M., 2007. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, *126*(1), p.014101.

Çakir, P., Cutivet, A., Resmini, M., Bui, B.T.S. and Haupt, K., 2013. Protein-Size Molecularly Imprinted Polymer Nanogels as Synthetic Antibodies, by Localized Polymerization with Multi-initiators. *Advanced Materials*, *25*(7), pp.1048-1051.

Cascella, M. and Vanni, S., 2015. Toward accurate coarse-graining approaches for protein and membrane simulations.

Castro, B., Ramirez, G., Rubio, M.F., Whitcombe, M.J., Vulfson, E.N., Vazquez-Duhalt, R. and Barzana, E., 1998. Molecular modelling as a tool for predicting ligand-receptor interactions in molecularly imprinted polymers used for the removal of organosulfur compounds from fuels, applying molecular modeling and computational chemistry. *AIChE J.*, pp.329-336.

Ceccon, A., Busato, M., Pérez Santero, S., D'Onofrio, M., Musiani, F., Giorgetti, A. and Assfalg, M., 2015. Transient Interactions of a Cytosolic Protein with Macromolecular and Vesicular Cosolutes: Unspecific and Specific Effects. *ChemBioChem*, *16*(18), pp.2633-2645.

Cellini, B., Bertoldi, M., Montioli, R., Paiardini, A. and Voltattorni, C.B., 2007. Human wild-type alanine: glyoxylate aminotransferase and its naturally occurring G82E variant: functional properties and physiological implications. *Biochemical Journal*, *408*(1), pp.39-50.

Cellini, B., Lorenzetto, A., Montioli, R., Oppici, E. and Voltattorni, C.B., 2010. Human liver peroxisomal alanine: glyoxylate aminotransferase: Different stability under chemical stress of the major allele, the minor allele, and its pathogenic G170R variant. *Biochimie*, *92*(12), pp.1801-1811.

Cellini, B., Montioli, R., Paiardini, A., Lorenzetto, A., Maset, F., Bellini, T., Oppici, E. and Voltattorni, C.B., 2010. Molecular defects of the glycine 41 variants of alanine glyoxylate aminotransferase associated with primary hyperoxaluria type I. *Proceedings of the National Academy of Sciences*, *107*(7), pp.2896-2901.

Cenci, L., Andreetto, E., Vestri, A., Bovi, M., Barozzi, M., Iacob, E., Busato, M., Castagna, A., Girelli, D. and Bossi, A.M., 2015. Surface plasmon resonance based on molecularly imprinted nanoparticles for the picomolar detection of the iron regulating hormone Hepcidin-25. Journal of nanobiotechnology, 13(1), p.1.

Cenci, L., Anesi, A., Busato, M., Guella, G. and Bossi, A.M., 2016. Molecularly imprinted polymers coupled to matrix assisted laser desorption ionization mass spectrometry for

femtomoles detection of cardiac troponin I peptides. *Journal of Molecular Recognition*, *29*(1), pp.41-50.

Chaudhuri, R., Lindwasser, O.W., Smith, W.J., Hurley, J.H. and Bonifacino, J.S., 2007. Downregulation of CD4 by human immunodeficiency virus type 1 Nef is dependent on clathrin and involves direct interaction of Nef with the AP2 clathrin adaptor. *Journal of virology*, *81*(8), pp.3877-3890.

Chen, H. and Zhou, H.X., 2005. Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. *Proteins: Structure, Function, and Bioinformatics*, *61*(1), pp.21-35.

Chen, J., Brooks, C.L. and Khandogin, J., 2008. Recent advances in implicit solvent-based methods for biomolecular simulations. *Current Opinion in Structural Biology*, *18*(2), pp.140-148.

Chianella, I., Lotierzo, M., Piletsky, S.A., Tothill, I.E., Chen, B., Karim, K. and Turner, A.P., 2002. Rational design of a polymer specific for microcystin-LR using a computational approach. *Analytical chemistry*, *74*(6), pp.1288-1293.

Chong, L.T., Duan, Y., Wang, L., Massova, I. and Kollman, P.A., 1999. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. *Proceedings of the National Academy of Sciences*, *96*(25), pp.14330-14335.

Chothia, C. and Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. The EMBO journal, 5(4), p.823.

Christen, M., Hünenberger, P.H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D.P., Heinz, T.N., Kastenholz, M.A., Kräutler, V., Oostenbrink, C. and Peter, C., 2005. The GROMOS software for biomolecular simulation: GROMOS05. Journal of computational chemistry, 26(16), pp.1719-1751.

Clark, M., Cramer, R.D. and Van Opdenbosch, N., 1989. Validation of the general purpose Tripos 5.2 force field. *Journal of Computational Chemistry*, *10*(8), pp.982-1012.

Cohen, G.B., Gandhi, R.T., Davis, D.M., Mandelboim, O., Chen, B.K., Strominger, J.L. and Baltimore, D., 1999. The selective downregulation of class I major histocompatibility complex proteins by HIV-1 protects HIV-infected cells from NK cells. *Immunity*, *10*(6), pp.661-671.

Cohen, G.B., Rangan, V.S., Chen, B.K., Smith, S. and Baltimore, D., 2000. The human thioesterase II protein binds to a site on HIV-1 Nef critical for CD4 down-regulation. *Journal of Biological Chemistry*, *275*(30), pp.23097-23105.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. and Kollman, P.A., 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, *117*(19), pp.5179-5197.

Cowen, T., Busato, M., Karim, K. and Piletsky, S.A., 2016. In Silico Synthesis of Synthetic Receptors: A Polymerization Algorithm. *Macromolecular Rapid Communications*.

Crowley, P.B. and Ubbink, M., 2003. Close encounters of the transient kind: protein interactions in the photosynthetic redox chain investigated by NMR spectroscopy. *Accounts of chemical research*, *36*(10), pp.723-730.

Crowley, P.B., Chow, E. and Papkovskaia, T., 2011. Protein Interactions in the Escherichia coli Cytosol: An Impediment toIn-Cell NMR Spectroscopy. *ChemBioChem*, *12*(7), pp.1043-1048.

Cyril Dominguez, Rolf Boelens and Alexandre M.J.J. Bonvin. HADDOCK: a protein-protein docking approach based on biochemical and/or biophysical information. J. Am. Chem. Soc. 125, 1731-1737 (2003).

da Silva, A.W.S. and Vranken, W.F., 2012. ACPYPE-Antechamber python parser interface. *BMC research notes*, *5*(1), p.1.

da Silva, J. and Hughes, A.L., 1998. Conservation of cytotoxic T lymphocyte (CTL) epitopes as a host strategy to constrain parasite adaptation: evidence from the nef gene of human immunodeficiency virus 1 (HIV-1). *Molecular biology and evolution*, *15*(10), pp.1259-1268.

Danpure, C.J., 1986. Peroxisomal alanine: glyoxylate aminotransferase and prenatal diagnosis of primary hyperoxaluria type 1. *The Lancet*, *328*(8516), p.1168.

de Vries, S.J. and Bonvin, A.M., 2011. CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PloS one*, *6*(3), p.e17695.

de Vries, S.J., van Dijk, A.D. and Bonvin, A.M., 2006. WHISCY: What information does surface conservation yield? Application to data-driven docking. *Proteins: Structure, Function, and Bioinformatics*, *63*(3), pp.479-489.

De Vries, S.J., van Dijk, A.D., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T. and Bonvin, A.M., 2007. HADDOCK versus HADDOCK: new features and performance of HADDOCK2. 0 on the CAPRI targets. *Proteins: structure, function, and bioinformatics*, *69*(4), pp.726-733.

Deacon, N.J., Tsykin, A., Solomon, A. and Smith, K., 1995. Genomic structure of an attenuated quasi species of HIV-1 from a blood transfusion donor and recipients. *Science*, *270*(5238), p.988.

Dindo, M., Montioli, R., Busato, M., Giorgetti, A., Cellini, B. and Voltattorni, C.B., 2016. Effects of interface mutations on the dimerization of alanine glyoxylate aminotransferase and implications in the mistargeting of the pathogenic variants F152I and I244T. Biochimie, 131, pp.137-148.

Dominguez, C., Boelens, R. and Bonvin, A.M., 2003. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, *125*(7), pp.1731-1737.

Donald, J.E., Kulp, D.W. and DeGrado, W.F., 2011. Salt bridges: geometrically specific, designable interactions. *Proteins: Structure, Function, and Bioinformatics*, *79*(3), pp.898-915.

Dong, W., Yan, M., Zhang, M., Liu, Z. and Li, Y., 2005. A computational and experimental investigation of the interaction between the template molecule and the functional monomer used in the molecularly imprinted polymer. *Analytica chimica acta*, *542*(2), pp.186-192..

Dourado, E., Herdes, C., Tassel, P.R.V. and Sarkisov, L., 2011. Molecular recognition effects in atomistic models of imprinted polymers. *International journal of molecular sciences*, *12*(8), pp.4781-4804.

Dupradeau, F.Y., Pigache, A., Zaffran, T., Savineau, C., Lelong, R., Grivel, N., Lelong, D., Rosanski, W. and Cieplak, P., 2010. The REd. Tools: Advances in RESP and ESP charge derivation and force field library building. *Physical Chemistry Chemical Physics*, *12*(28), pp.7821-7839.

Eason, M.G., Jacinto, M.T., Theiss, C.T. and Liggett, S.B., 1994. The palmitoylated cysteine of the cytoplasmic tail of alpha 2A-adrenergic receptors confers subtype-specific agonist-promoted downregulation. *Proceedings of the National Academy of Sciences*, *91*(23), pp.11178-11182.

Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics, 14(9), pp.755-763.

Eliseo, T., Ragona, L., Catalano, M., Assfalg, M., Paci, M., Zetta, L., Molinari, H. and Cicero, D.O., 2007. Structural and dynamic determinants of ligand binding in the ternary complex of chicken liver bile acid binding protein with two bile salts revealed by NMR. *Biochemistry*, *46*(44), pp.12557-12567.

Eswar, N., Eramian, D., Webb, B., Shen, M.Y. and Sali, A., 2008. Protein structure modeling with MODELLER. *Structural proteomics: high-throughput methods*, pp.145-159.

Eswar, N., Marti-Renom, M. A., Webb, B., Madhusudhan, M. S., Eramian, D., Shen, M., et al. (2006) Comparative Protein Structure Modeling With MODELLER. Current Protocols in Bioinformatics, John Wiley& Sons, Inc., Supplement 15, 5.6.1-5.6.30.

Fargue, S., Lewin, J., Rumsby, G. and Danpure, C.J., 2013. Four of the most common mutations in primary hyperoxaluria type 1 unmask the cryptic mitochondrial targeting sequence of alanine: glyoxylate aminotransferase encoded by the polymorphic minor allele. *Journal of Biological Chemistry*, *288*(4), pp.2475-2484.

Fiser, A., 2010. Template-based protein structure modeling. Computational Biology, pp.73-94.

Fiser, A., Do, R.K.G. and Šali, A., 2000. Modeling of loops in protein structures. Protein science, 9(09), pp.1753-1773.

Forli, S., Huey, R., Pique, M.E., Sanner, M.F., Goodsell, D.S. and Olson, A.J., 2016. Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nature protocols*, *11*(5), pp.905-919.

Frenkel, D. and Smit, B., 2002. Understanding molecular simulation: from algorithms to applications. *Computational sciences series*, *1*, pp.1-638.

Fu, Q., Sanbe, H., Kagawa, C., Kunimoto, K.K. and Haginaka, J., 2003. Uniformly sized molecularly imprinted polymer for (S)-nilvadipine. Comparison of chiral recognition ability with HPLC chiral stationary phases based on a protein. *Analytical chemistry*, *75*(2), pp.191-198.

G.C.P van Zundert, J.P.G.L.M. Rodrigues, M. Trellet, C. Schmitz, P.L. Kastritis, E. Karaca, A.S.J. Melquiond, M. van Dijk, S.J. de Vries and A.M.J.J. Bonvin. "The HADDOCK2.2 webserver: User-friendly integrative modeling of biomolecular complexes." J. Mol. Biol., 428, 720-725 (2016).

Gierasch, L.M. and Gershenson, A., 2009. Post-reductionist protein science, or putting Humpty Dumpty back together again. *Nature chemical biology*, *5*(11), pp.774-777.

Golker, K. and Nicholls, I.A., 2016. The effect of crosslinking density on molecularly imprinted polymer morphology and recognition. *European Polymer Journal*, *75*, pp.423-430.

Golker, K., Karlsson, B.C., Rosengren, A.M. and Nicholls, I.A., 2014. A functional monomer is not enough: principal component analysis of the influence of template complexation in pre-polymerization mixtures on imprinted polymer recognition and morphology. *International journal of molecular sciences*, *15*(11), pp.20572-20584.

Gossner, C.M.E., Schlundt, J., Embarek, P.B., Hird, S., Lo-Fo-Wong, D., Beltran, J.J.O., Teoh, K.N. and Tritscher, A., 2009. The melamine incident: implications for international food and feed safety. *Environmental health perspectives*, pp.1803-1808.

Gould, H., Tobochnik, J. and Christian, W., 1988. *An introduction to computer simulation methods* (Vol. 1). New York: Addison-Wesley.

Graziani, A., Galimi, F., Medico, E., Cottone, E., Gramaglia, D., Boccaccio, C. and Comoglio, P.M., 1996. The HIV-1 nef protein interferes with phosphatidylinositol 3-kinase activation 1. *Journal of Biological Chemistry*, *271*(12), pp.6590-6593.

Guex, N., Peitsch, M.C. and Schwede, T., 2009. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. Electrophoresis, 30(S1), pp.S162-S173.

Halgren, T.A., 1996. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *Journal of Computational Chemistry*, *17*(5-6), pp.520-552.

Harris, R., Olson, A.J. and Goodsell, D.S., 2008. Automated prediction of ligand-binding sites in proteins. *Proteins: Structure, Function, and Bioinformatics*, *70*(4), pp.1506-1517.

Henikoff, S. and Henikoff, J.G., 1994. Protein family classification based on searching a database of blocks. Genomics, 19(1), pp.97-107.

Hess, B., Bekker, H., Berendsen, H.J. and Fraaije, J.G., 1997. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry*, *18*(12), pp.1463-1472.

Hess, B., Kutzner, C., Van Der Spoel, D. and Lindahl, E., 2008. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation*, *4*(3), pp.435-447.

Hockney, R. and Eastwood, J., 1988. Computer Simulations using Particles, Adam Higler Bristol New York.

Hofmann, A.F., 2011. Enterohepatic circulation of bile acids. *Comprehensive Physiology*.

Hooft, R.W., Vriend, G., Sander, C. and Abola, E.E., 1996. Errors in protein structures. Nature, 381(6580), pp.272-272.

Hoover, W.G., 1985. Canonical dynamics: equilibrium phase-space distributions. *Physical review A*, *31*(3), p.1695.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. and Simmerling, C., 2006. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, *65*(3), pp.712-725.

Hug, S., 2013. Classical molecular dynamics in a nutshell. *Biomolecular Simulations: Methods and Protocols*, pp.127-152.

Hunt, M.C., Solaas, K., Kase, B.F. and Alexson, S.E., 2002. Characterization of an acyl-CoA thioesterase that functions as a major regulator of peroxisomal lipid metabolism. *Journal of Biological Chemistry*, *277*(2), pp.1128-1138.

Iafrate, A.J., Bronson, S. and Skowronski, J., 1997. Separable functions of Nef disrupt two aspects of T cell receptor machinery: CD4 expression and CD3 signaling. *The EMBO Journal*, *16*(4), pp.673-684.

Irwin, J.J., Sterling, T., Mysinger, M.M., Bolstad, E.S. and Coleman, R.G., 2012. ZINC: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, *52*(7), pp.1757-1768.

Janin, J., Bahadur, R.P. and Chakrabarti, P., 2008. Protein–protein interaction and quaternary structure. *Quarterly reviews of biophysics*, *41*(02), pp.133-180.

Janin, J., Henrick, K., Moult, J., Eyck, L.T., Sternberg, M.J., Vajda, S., Vakser, I. and Wodak, S.J., 2003. CAPRI: a critical assessment of predicted interactions. *Proteins: Structure, Function, and Bioinformatics*, *52*(1), pp.2-9.

Jones, J.M., Nau, K., Geraghty, M.T., Erdmann, R. and Gould, S.J., 1999. Identification of peroxisomal acyl-CoA thioesterases in yeast and humans. *Journal of Biological Chemistry*, *274*(14), pp.9216-9223.

Jorgensen, W.L. and Tirado-Rives, J., 1988. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, *110*(6), pp.1657-1666.

Jorgensen, W.L., Maxwell, D.S. and Tirado-Rives, J., 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, *118*(45), pp.11225-11236.

Kabsch, W. and Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), pp.2577-2637.

Karim, K., Breton, F., Rouillon, R., Piletska, E.V., Guerreiro, A., Chianella, I. and Piletsky, S.A., 2005. How to find effective functional monomers for effective molecularly imprinted polymers?. *Advanced drug delivery reviews*, *57*(12), pp.1795-1808.

Karlsson, B.C., O'Mahony, J., Karlsson, J.G., Bengtsson, H., Eriksson, L.A. and Nicholls, I.A., 2009. Structure and dynamics of monomer− template complexation: an explanation for molecularly imprinted polymer recognition site heterogeneity. *Journal of the American Chemical Society*, *131*(37), pp.13297-13304.

Kass, I., Hoke, D.E., Costa, M.G., Reboul, C.F., Porebski, B.T., Cowieson, N.P., Leh, H., Pennacchietti, E., McCoey, J., Kleifeld, O. and Voltattorni, C.B., 2014. Cofactor-dependent conformational heterogeneity of GAD65 and its role in autoimmunity and neurotransmitter homeostasis. *Proceedings of the National Academy of Sciences*, *111*(25), pp.E2524-E2529.

Kastritis, P.L. and Bonvin, A.M., 2013. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface*, *10*(79), p.20120835.

Katz, A. and Davis, M.E., 1999. Investigations into the mechanisms of molecular recognition with imprinted polymers. *Macromolecules*, *32*(12), pp.4113-4121.

Keskin, O., Gursoy, A., Ma, B. and Nussinov, R., 2008. Principles of protein-protein interactions: what are the preferred ways for proteins to interact?. *Chemical reviews*, *108*(4), pp.1225-1244.

Kestier, H.W., Ringler, D.J., Mori, K., Panicali, D.L., Sehgal, P.K., Daniel, M.D. and Desrosiers, R.C., 1991. Importance of the nef gene for maintenance of high virus loads and for development of AIDS. *Cell*, *65*(4), pp.651-662.

Kiefer, F., Arnold, K., Künzli, M., Bordoli, L. and Schwede, T., 2009. The SWISS-MODEL Repository and associated resources. Nucleic acids research, 37(suppl 1), pp.D387-D392.

Kienzle, N., Freund, J., Kalbitzer, H.R. and MUELLER-LANTZSCH, N., 1993. Oligomerization of the Nef protein from human immunodeficiency virus (HIV) type 1. *European Journal of Biochemistry*, *214*(2), pp.451-457.

Kirchhoff, F., Greenough, T.C., Brettler, D.B., Sullivan, J.L. and Desrosiers, R.C., 1995. Absence of intact nef sequences in a long-term survivor with nonprogressive HIV-1 infection. *New England Journal of Medicine*, *332*(4), pp.228-232.

Kramer, W., Girbig, F., Gutjahr, U., Kowalewski, S., Jouvenal, K., Müller, G., Tripier, D. and Wess, G., 1993. Intestinal bile acid absorption. Na (+)-dependent bile acid transport activity in rabbit small intestine correlates with the coexpression of an integral 93-kDa and a peripheral 14-kDa bile acid-binding membrane protein along the duodenum-ileum axis. *Journal of Biological Chemistry*, *268*(24), pp.18035-18046.

Krieger, E., Joo, K., Lee, J., Lee, J., Raman, S., Thompson, J., Tyka, M., Baker, D. and Karplus, K., 2009. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: four approaches that performed well in CASP8. Proteins: Structure, Function, and Bioinformatics, 77(S9), pp.114-122.

Krogh, A., Brown, M., Mian, I.S., Sjölander, K. and Haussler, D., 1994. Hidden Markov models in computational biology: Applications to protein modeling. Journal of molecular biology, 235(5), pp.1501-1531.

Kufareva, I., Budagyan, L., Raush, E., Totrov, M. and Abagyan, R., 2007. PIER: protein interface recognition for structural proteomics. *Proteins: Structure, Function, and Bioinformatics*, *67*(2), pp.400-417.

Kufareva, I., Katritch, V., Stevens, R.C. and Abagyan, R., 2014. Advances in GPCR modeling evaluated by the GPCR Dock 2013 assessment: meeting new challenges. *Structure*, *22*(8), pp.1120-1139.

Kumar, A., Rajendran, V., Sethumadhavan, R. and Purohit, R., 2013. Molecular dynamic simulation reveals damaging impact of RAC1 F28L mutation in the switch I region. *PloS one*, *8*(10), p.e77453.

Kurz, M., Brachvogel, V., Matter, H., Stengelin, S., Thüring, H. and Kramer, W., 2003. Insights into the bile acid transportation system: The human ileal lipid-binding protein-cholyltaurine complex and its comparison with homologous structures. *Proteins: Structure, Function, and Bioinformatics*, *50*(2), pp.312-328.

Kuznetsova, I.M., Turoverov, K.K. and Uversky, V.N., 2014. What macromolecular crowding can do to a protein. *International journal of molecular sciences*, *15*(12), pp.23090-23140.

Lanza, F. and Sellergren, B., 1999. Method for synthesis and screening of large groups of molecularly imprinted polymers. *Analytical chemistry*, *71*(11), pp.2092-2096.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M., 1993. PROCHECK: a program to check the stereochemical quality of protein structures. Journal of applied crystallography, 26(2), pp.283-291.

Leach, A.R., 2001. *Molecular modelling: principles and applications*. Pearson education.

Levitt, M., 1992. Accurate modeling of protein conformation by automatic segment matching. Journal of molecular biology, 226(2), pp.507-533.

Levitt, M., 2001. The birth of computational structural biology. *Nature Structural & Molecular Biology*, *8*(5), pp.392-393.

Li, J., Derewenda, U., Dauter, Z., Smith, S. and Derewenda, Z.S., 2000. Crystal structure of the Escherichia coli thioesterase II, a homolog of the human Nef binding enzyme. *Nature Structural & Molecular Biology*, *7*(7), pp.555-559.

Li, Y., Xu, J. and Sun, C., 2015. Chemical sensors and biosensors for the detection of melamine. *RSC Advances*, *5*(2), pp.1125-1147.

Liang, S., Zhang, C., Liu, S. and Zhou, Y., 2006. Protein binding site prediction using an empirical scoring function. *Nucleic acids research*, *34*(13), pp.3698-3707.

Lindahl, E.R., 2008. Molecular dynamics simulations. *Molecular modeling of proteins*, pp.3-23.

Lipman, D.J. and Pearson, W.R., 1985. Rapid and sensitive protein similarity searches. Science, 227(4693), pp.1435-1441.

Liu, L.X., Heveker, N., Fackler, O.T., Arold, S., Le Gall, S., Janvier, K., Peterlin, B.M., Dumas, C., Schwartz, O., Benichou, S. and Benarous, R., 2000. Mutation of a conserved residue (D123) required for oligomerization of human immunodeficiency virus type 1 Nef protein abolishes interaction with human thioesterase and results in impairment of Nef biological functions. *Journal of virology*, *74*(11), pp.5310-5319.

Liu, L.X., Margottin, F., Le Gall, S., Schwartz, O., Selig, L., Benarous, R. and Benichou, S., 1997. Binding of HIV-1 Nef to a novel thioesterase enzyme correlates with Nef-mediated CD4 down-regulation. *Journal of Biological Chemistry*, *272*(21), pp.13779-13785.

Lucchiari-Hartz, M., Van Endert, P.M., Lauvau, G., Maier, R., Meyerhans, A., Mann, D., Eichmann, K. and Niedermann, G., 2000. Cytotoxic T Lymphocyte Epitopes of HIV-1 Nef Generation of Multiple Definitive Major Histocompatibility Complex Class I Ligands by Proteasomes. *The Journal of experimental medicine*, *191*(2), pp.239-252.

Lv, Y., Lin, Z., Tan, T., Feng, W., Qin, P. and Li, C., 2008. Application of molecular dynamics modeling for the prediction of selective adsorption properties of dimethoate imprinting polymer. *Sensors and Actuators B: Chemical*, *133*(1), pp.15-23.

Lynch, J.M. and Barbano, D.M., 1999. Kjeldahl nitrogen analysis as a reference method for protein determination in dairy products. *Journal-aoac international*, *82*, pp.1389-1398.

Magner, A., Szpankowski, W. and Kihara, D., 2015. On the origin of protein superfamilies and superfolds. Scientific reports, 5.

Makoto K, Toshifumi T, Takashi T, Hiroyuki A (2004): Molecular Imprinting:From Fundamentals to Applications. ed. Wiley VCH.

Malde, A.K., Zuo, L., Breeze, M., Stroet, M., Poger, D., Nair, P.C., Oostenbrink, C. and Mark, A.E., 2011. An automated force field topology builder (ATB) and repository: version 1.0. *Journal of chemical theory and computation*, *7*(12), pp.4026-4037.

Mariani, V., Kiefer, F., Schmidt, T., Haas, J. and Schwede, T., 2011. Assessment of template based protein structure predictions in CASP9. Proteins: Structure, Function, and Bioinformatics, 79(S10), pp.37-58.

Mark, P. and Nilsson, L., 2001. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *The Journal of Physical Chemistry A*, *105*(43), pp.9954-9960.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R. and Sander, C., 2011. Protein 3D structure computed from evolutionary sequence variation. *PloS one*, *6*(12), p.e28766.

Marrink, S.J., De Vries, A.H. and Mark, A.E., 2004. Coarse grained model for semiquantitative lipid simulations. *The Journal of Physical Chemistry B*, *108*(2), pp.750-760.

Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P. and De Vries, A.H., 2007. The MARTINI force field: coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, *111*(27), pp.7812-7824.

Marti-Renom, M.A., Madhusudhan, M.S. and Sali, A., 2004. Alignment of protein sequences by their profiles. Protein Science, 13(4), pp.1071-1087.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sánchez, R., Melo, F. and Šali, A., 2000. Comparative protein structure modeling of genes and genomes. Annual review of biophysics and biomolecular structure, 29(1), pp.291-325.

Meier, A. and Söding, J., 2015. Automatic prediction of protein 3D structures by probabilistic multi-template homology modeling. PLoS Comput Biol, 11(10), p.e1004343.

Melo, F. and Feytmans, E., 1998. Assessing protein structures with a non-local atomic interaction energy. Journal of molecular biology, 277(5), pp.1141-1152.

Miller, M.D., Warmerdam, M.T., Gaston, I., Greene, W.C. and Feinberg, M.B., 1994. The human immunodeficiency virus-1 nef gene product: a positive factor for viral infection and replication in primary lymphocytes and macrophages. *The Journal of experimental medicine*, *179*(1), pp.101-113.

Minton, A.P., 2013. Quantitative assessment of the relative contributions of steric repulsion and chemical interactions to macromolecular crowding. *Biopolymers*, *99*(4), pp.239-244.

Miura, S., Kasuya-Arai, I., Mori, H., Miyazawa, S., Osumi, T., Hashimoto, T. and Fujiki, Y., 1992. Carboxyl-terminal consensus Ser-Lys-Leu-related tripeptide of peroxisomal proteins functions in vitro as a minimal peroxisome-targeting signal. *Journal of Biological Chemistry*, *267*(20), pp.14405-14411.

Molinelli, A., O'Mahony, J., Nolan, K., Smyth, M.R., Jakusch, M. and Mizaikoff, B., 2005. Analyzing the mechanisms of selectivity in biomimetic self-assemblies via IR and NMR spectroscopy of prepolymerization solutions and molecular dynamics simulations. *Analytical chemistry*, *77*(16), pp.5196-5204.

Monti, S., Cappelli, C., Bronco, S., Giusti, P. and Ciardelli, G., 2006. Towards the design of highly selective recognition sites into molecular imprinting polymers: a computational approach. *Biosensors and Bioelectronics*, *22*(1), pp.153-163.

Monticelli, L., Kandasamy, S.K., Periole, X., Larson, R.G., Tieleman, D.P. and Marrink, S.J., 2008. The MARTINI coarse-grained force field: extension to proteins. *Journal of chemical theory and computation*, *4*(5), pp.819-834.

Montioli, R., Oppici, E., Dindo, M., Roncador, A., Gotte, G., Cellini, B. and Voltattorni, C.B., 2015. Misfolding caused by the pathogenic mutation G47R on the minor allele of alanine: glyoxylate aminotransferase and chaperoning activity of pyridoxine. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1854*(10), pp.1280-1289.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. and Olson, A. J. (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity. J. Computational Chemistry 2009, 16: 2785-91.

Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K. and Olson, A.J., 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of computational chemistry, 19(14), pp.1639-1662.

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S. and Olson, A.J., 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, *30*(16), pp.2785-2791.

Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, *247*(4), pp.536-540.

Musiani, F., Bertoša, B., Magistrato, A., Zambelli, B., Turano, P., Losasso, V., Micheletti, C., Ciurli, S. and Carloni, P., 2010. Computational Study of the DNA-Binding Protein Helicobacter pylori NikR: The Role of Ni2+ 2 2 Francesco Musiani and Branimir Bertoša contributed equally to the simulations presented here. *Journal of chemical theory and computation*, *6*(11), pp.3503-3515.

Najmanovich, R., Kurbatova, N. and Thornton, J., 2008. Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics*, *24*(16), pp.i105-i111.

National Toxicology Program, 1983. NTP Carcinogenesis Bioassay of Melamine (CAS No. 108-78-1) in F344/N Rats and B6C3F1 Mice (Feed Study). *National Toxicology Program technical report series*, *245*, p.1.

Nauchitel', V.V., 1981. Energy distribution function for the NVT canonical ensemble. *Molecular Physics*, *42*(5), pp.1259-1265.

Neuvirth, H., Raz, R. and Schreiber, G., 2004. ProMate: a structure based prediction program to identify the location of protein–protein binding sites. *Journal of molecular biology*, *338*(1), pp.181-199.

Nicholls, I.A., 1998. Towards the rational design of molecularly imprinted polymers. *Journal of Molecular Recognition*, *11*(1-6), pp.79-82.

Nicholls, I.A., Andersson, H.S., Charlton, C., Henschel, H., Karlsson, B.C., Karlsson, J.G., O'Mahony, J., Rosengren, A.M., Rosengren, K.J. and Wikman, S., 2009. Theoretical and computational strategies for rational molecularly imprinted polymer design. *Biosensors and Bioelectronics*, *25*(3), pp.543-552

Nooren, I.M. and Thornton, J.M., 2003. Diversity of protein–protein interactions. *The EMBO journal*, *22*(14), pp.3486-3492.

Nosé, S. and Klein, M.L., 1983. Constant pressure molecular dynamics for molecular systems. *Molecular Physics*, *50*(5), pp.1055-1076.

Nosé, S., 1984. A molecular dynamics method for simulations in the canonical ensemble. *Molecular physics*, *52*(2), pp.255-268.

Nugent, T. and Jones, D.T., 2012. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proceedings of the National Academy of Sciences*, *109*(24), pp.E1540-E1547.

Nussinov, R. and Jang, H., 2014. Dynamic multiprotein assemblies shape the spatial structure of cell signaling. *Progress in biophysics and molecular biology*, *116*(2), pp.158-164.

Olsson, G.D., Karlsson, B.C., Shoravi, S., Wiklander, J.G. and Nicholls, I.A., 2012. Mechanisms underlying molecularly imprinted polymer molecular memory and the role of crosslinker: Resolving debate on the nature of template recognition in phenylalanine anilide imprinted polymers. *Journal of Molecular Recognition*, *25*(2), pp.69-73.

Oostenbrink, C., Villa, A., Mark, A.E. and Van Gunsteren, W.F., 2004. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of computational chemistry*, *25*(13), pp.1656-1676.

Oppici, E., Roncador, A., Montioli, R., Bianconi, S. and Cellini, B., 2013. Gly161 mutations associated with Primary Hyperoxaluria Type I induce the cytosolic aggregation and the intracellular degradation of the apo-form of alanine: glyoxylate aminotransferase. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1832*(12), pp.2277-2288.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M., 1997. CATH–a hierarchic classification of protein domain structures. Structure, 5(8), pp.1093-1109.

Ozbabacan, S.E.A., Engin, H.B., Gursoy, A. and Keskin, O., 2011. Transient protein–protein interactions. *Protein Engineering Design and Selection*, *24*(9), pp.635-648.

Palomba, D., Cavasotto, C., 2015. Protein structure modelling in drug design. In: Cavasotto, C. (Ed.), In Silico Drug Discovery and Design: Theory, Methods, Challenges, and Applications. CRC Press, pp. 215–247

Panuwet, P., Wade, E.L., Nguyen, J.V., Montesano, M.A., Needham, L.L. and Barr, D.B., 2010. Quantification of cyanuric acid residue in human urine using high performance liquid chromatography–tandem mass spectrometry. *Journal of Chromatography B*, *878*(28), pp.2916-2922.

Parrinello, M. and Rahman, A., 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, *52*(12), pp.7182-7190.

Paulechka, E., Kroenlein, K., Kazakov, A. and Frenkel, M., 2012. A systematic approach for development of an OPLS-Like force field and its application to hydrofluorocarbons. *The Journal of Physical Chemistry B*, *116*(49), pp.14389-14397.

Pavel, D. and Lagowski, J., 2005. Computationally designed monomers and polymers for molecular imprinting of theophylline—part II. *Polymer*, *46*(18), pp.7543-7556.

Pei, J., Kim, B.H., Tang, M. and Grishin, N.V., 2007. PROMALS web server for accurate multiple protein sequence alignments. *Nucleic acids research*, *35*(suppl 2), pp.W649-W652.

Periole, X., Cavalli, M., Marrink, S.J. and Ceruso, M.A., 2009. Combining an elastic network with a coarse-grained molecular force field: structure, dynamics, and intermolecular recognition. *Journal of Chemical Theory and Computation*, *5*(9), pp.2531-2543.

Perkins, J.R., Diboun, I., Dessailly, B.H., Lees, J.G. and Orengo, C., 2010. Transient protein-protein interactions: structural, functional, and network properties. *Structure*, *18*(10), pp.1233-1243.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E., 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, *25*(13), pp.1605-1612.

Pey, A.L., Salido, E. and Sanchez-Ruiz, J.M., 2011. Role of low native state kinetic stability and interaction of partially unfolded states with molecular chaperones in the mitochondrial protein mistargeting associated with primary hyperoxaluria. *Amino acids*, *41*(5), pp.1233-1245.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L. and Schulten, K., 2005. Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, *26*(16), pp.1781-1802.

Piccoli, S., Musiani, F. and Giorgetti, A., 2014. Dynamic characterization and substrate binding of cis-2, 3-dihydrobiphenyl-2, 3-diol dehydrogenase—an enzyme used in bioremediation. *Journal of molecular modeling*, *20*(12), pp.1-6.

Piletsky, S.A., Karim, K., Piletska, E.V., Day, C.J., Freebairn, K.W., Legge, C. and Turner, A.P.F., 2001. Recognition of ephedrine enantiomers by molecularly imprinted polymers designed using a computational approach. *Analyst*, *126*(10), pp.1826-1830.

Porollo, A. and Meller, J., 2007. Prediction-based fingerprints of protein–protein interactions. *Proteins: structure, function, and bioinformatics*, *66*(3), pp.630-645.

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M.R., Smith, J.C., Kasson, P.M., van der Spoel, D. and Hess, B., 2013. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, p.btt055.

Purdue, P.E., Allsop, J., Isaya, G., Rosenberg, L.E. and Danpure, C.J., 1991. Mistargeting of peroxisomal L-alanine: glyoxylate aminotransferase to mitochondria in primary hyperoxaluria patients depends upon activation of a cryptic mitochondrial targeting sequence by a point mutation. *Proceedings of the National Academy of Sciences*, *88*(23), pp.10900-10904.

Qiao, X., He, B., Chiu, A., Knowles, D.M., Chadburn, A. and Cerutti, A., 2006. Human immunodeficiency virus 1 Nef suppresses CD40-dependent immunoglobulin class switching in bystander B cells. *Nature immunology*, *7*(3), pp.302-310.

Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A. and Pandey, G., 2013. A large-scale evaluation of computational protein function prediction. *Nature methods*, *10*(3), pp.221-227.

Rajith, B., Chakraborty, C., NagaSundaram, N., Ali, S.K. and Zhu, H., 2014. Structural signature of the G719S-T790M double mutation in the EGFR kinase domain and its response to inhibitors. *Scientific reports*, *4*.

Remmert, M., Biegert, A., Hauser, A. and Söding, J., 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods, 9(2), pp.173-175.

Robert, X. and Gouet, P., 2014. Deciphering key features in protein structures with the new ENDscript server. *Nucleic acids research*, *42*(W1), pp.W320-W324.

Rodrigues, J.P.G.L.M., Melquiond, A.S.J., Karaca, E., Trellet, M., Dijk, M., Zundert, G.C.P., Schmitz, C., Vries, S.J., Bordogna, A., Bonati, L. and Kastritis, P.L., 2013. Defining the limits of homology modeling in information-driven protein docking. Proteins: Structure, Function, and Bioinformatics, 81(12), pp.2119-2128.

Rost, B., 1999. Twilight zone of protein sequence alignments. Protein engineering, 12(2), pp.85-94.

Rothman, J.E., 1994. Mechanisms of intracellular protein transport.

Rovina, K. and Siddiquee, S., 2015. A review of recent advances in melamine detection techniques. *Journal of Food Composition and Analysis*, *43*, pp.25-38.

Sacchettini, J.C., Gordon, J.I. and Banaszak, L.J., 1988. The structure of crystalline Escherichia coli-derived rat intestinal fatty acid-binding protein at 2.5-A resolution. *Journal of Biological Chemistry*, *263*(12), pp.5815-5819.

Sali, A. and Blundell, T., 1994. Comparative protein modelling by satisfaction of spatial restraints. Protein structure by distance analysis, 64, p.C86.

Salido, E., Pey, A.L., Rodriguez, R. and Lorenzo, V., 2012. Primary hyperoxalurias: disorders of glyoxylate detoxification. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1822*(9), pp.1453-1464.

Salomon-Ferrer, R., Case, D.A. and Walker, R.C., 2013. An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *3*(2), pp.198-210.

Salvi, R., Garbuglia, A.R., Di Caro, A., Pulciani, S., Montella, F. and Benedetto, A., 1998. Grossly defective nef gene sequences in a human immunodeficiency virus type 1-seropositive long-term nonprogressor. *Journal of virology*, *72*(5), pp.3646-3657.

Samish, I., 2009. Search and sampling in structural bioinformatics. *Structural Bioinformatics*, pp.207-236.

Samish, I., Bourne, P.E. and Najmanovich, R.J., 2015. Achievements and challenges in structural bioinformatics and computational biophysics. *Bioinformatics*, *31*(1), pp.146-150.

Samish, I., MacDermaid, C.M., Perez-Aguilar, J.M. and Saven, J.G., 2011. Theoretical and computational protein design. *Annual review of physical chemistry*, *62*, pp.129-149.

Sandal M., 2016. "Finding common structural traits of GPCRs by computational molecular biology approaches". (PhD thesis)

Sander, C. and Schneider, R., 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins: Structure, Function, and Bioinformatics, 9(1), pp.56-68.

Sanfelice, D., Adrover, M., Martorell, G., Pastore, A. and Temussi, P.A., 2012. Crowding versus molecular seeding: NMR studies of protein aggregation in hen egg white. *Journal of Physics: Condensed Matter*, *24*(24), p.244107.

Schaap, F.G., Van der Vusse, G.J. and Glatz, J.F., 2002. Evolution of the family of intracellular lipid binding proteins in vertebrates. In *Cellular Lipid Binding Proteins* (pp. 69-77). Springer US.

Schauperl, M. and Lewis, D.W., 2015. Probing the Structural and Binding Mechanism Heterogeneity of Molecularly Imprinted Polymers. *The Journal of Physical Chemistry B*, *119*(2), pp.563-571.

Schlegel, B., Sippl, W. and Höltje, H.D., 2005. Molecular dynamics simulations of bovine rhodopsin: influence of protonation states and different membrane-mimicking environments. Journal of molecular modeling, 12(1), pp.49-64.

SchuÈttelkopf, A.W. and Van Aalten, D.M., 2004. PRODRG: a tool for high-throughput crystallography of protein–ligand complexes. *Acta Crystallographica Section D: Biological Crystallography*, *60*(8), pp.1355-1363.

Schwartz, O., Maréchal, V., Friguet, B., Arenzana-Seisdedos, F. and Heard, J.M., 1998. Antiviral activity of the proteasome on incoming human immunodeficiency virus type 1. *Journal of virology*, *72*(5), pp.3845-3850.

Sellergren, B., 1997. Noncovalent molecular imprinting: antibody-like molecular recognition in polymeric network materials. *TrAC Trends in Analytical Chemistry*, *16*(6), pp.310-320.

Serena, M., Giorgetti, A., Busato, M., Gasparini, F., Diani, E., Romanelli, M.G. and Zipeto, D., 2016. Molecular characterization of HIV-1 Nef and ACOT8 interaction: insights from in silico structural predictions and in vitro functional assays. *Scientific reports*, *6*.

Shaw, D.E., Dror, R.O., Salmon, J.K., Grossman, J.P., Mackenzie, K.M., Bank, J.A., Young, C., Deneroff, M.M., Batson, B., Bowers, K.J. and Chow, E., 2009, November. Millisecond-scale molecular dynamics simulations on Anton. In *Proceedings of the conference on high performance computing networking, storage and analysis* (pp. 1-11). IEEE.

Shelley, J.C., Shelley, M.Y., Reeder, R.C., Bandyopadhyay, S. and Klein, M.L., 2001. A coarse grain model for phospholipid simulations. The Journal of Physical Chemistry B, 105(19), pp.4464-4470.

Shen, M.Y. and Sali, A., 2006. Statistical potential for assessment and prediction of protein structures. Protein science, 15(11), pp.2507-2524.

Sillitoe, I., Cuff, A.L., Dessailly, B.H., Dawson, N.L., Furnham, N., Lee, D., Lees, J.G., Lewis, T.E., Studer, R.A., Rentzsch, R. and Yeats, C., 2012. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic acids research*, p.gks1211.

Skinner, C.G., Thomas, J.D. and Osterloh, J.D., 2010. Melamine toxicity. *Journal of Medical Toxicology*, *6*(1), pp.50-55

Skowronski, J., Parks, D. and Mariani, R., 1993. Altered T cell activation and development in transgenic mice expressing the HIV-1 nef gene. *The EMBO journal*, *12*(2), p.703.

Sobiech, M., Żołek, T., Luliński, P. and Maciejewska, D., 2014. A computational exploration of imprinted polymer affinity based on voriconazole metabolites. *Analyst*, *139*(7), pp.1779-1788.

Söding, J., 2005. Protein homology detection by HMM–HMM comparison. Bioinformatics, 21(7), pp.951-960.

Söding, J., Biegert, A. and Lupas, A.N., 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, *33*(suppl 2), pp.W244-W248.

Song, Y., DiMaio, F., Wang, R.Y.R., Kim, D., Miles, C., Brunette, T.J., Thompson, J. and Baker, D., 2013. High-resolution comparative modeling with RosettaCM. Structure, 21(10), pp.1735-1742.

Spina, C.A., Kwoh, T.J., Chowers, M.Y., Guatelli, J.C. and Richman, D.D., 1994. The importance of nef in the induction of human immunodeficiency virus type 1 replication from primary quiescent CD4 lymphocytes. *The Journal of experimental medicine*, *179*(1), pp.115-123.

Srinivasan, N. and Blundell, T.L., 1993. An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. Protein engineering, 6(5), pp.501-512.

Subrahmanyam, S., Karim, K. and Piletsky, S.A., 2012. Computational approaches in the design of synthetic receptors. In *Designing receptors for the next generation of biosensors* (pp. 131-165). Springer Berlin Heidelberg.

Subrahmanyam, S., Piletsky, S.A., Piletska, E.V., Chen, B., Karim, K. and Turner, A.P., 2001. 'Bite-and-Switch'approach using computationally designed molecularly imprinted polymers for sensing of creatinine. *Biosensors and Bioelectronics*, *16*(9), pp.631-637.

Sutcliffe, M.J., Haneef, I., Carney, D. and Blundell, T.L., 1987. Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Engineering, 1(5), pp.377-384.

Swingler, S., Mann, A., Jacque, J.M., Brichacek, B., Sasseville, V.G., Williams, K., Lackner, A.A., Janoff, E.N., Wang, R., Fisher, D. and Stevenson, M., 1999. HIV-1 Nef mediates lymphocyte chemotaxis and activation by infected macrophages. *Nature medicine*, *5*(9), pp.997-1003.

Swope, W.C., Andersen, H.C., Berens, P.H. and Wilson, K.R., 1982. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, *76*(1), pp.637-649.

Takeuchi, T., Fukuma, D. and Matsui, J., 1999. Combinatorial molecular imprinting: an approach to synthetic polymer receptors. *Analytical chemistry*, *71*(2), pp.285-290.

Touw, W.G., Baakman, C., Black, J., te Beek, T.A., Krieger, E., Joosten, R.P. and Vriend, G., 2015. A series of PDB-related databanks for everyday needs. *Nucleic acids research*, *43*(D1), pp.D364-D368.

Tramontano, A. and Lesk, A.M., 2006. Protein structure prediction. John Wiley and Sons, Inc, Weinheim.

Tramontano, A., Cozzetto, D., Giorgetti, A. and Raimondo, D., 2008. The assessment of methods for protein structure prediction. Protein Structure Prediction, pp.43-57.

Trott, O. and Olson, A.J., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, *31*(2), pp.455-
461.

Van Aalten, D.M., Bywater, R., Findlay, J.B., Hendlich, M., Hooft, R.W. and Vriend, G., 1996. PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. *Journal of computer-aided molecular design*, *10*(3), pp.255-262.

van den Berg, B., Ellis, R.J. and Dobson, C.M., 1999. Effects of macromolecular crowding on protein folding and aggregation. *The EMBO journal*, *18*(24), pp.6927-6933.

Venkatasami, G. and Sowa, J.R., 2010. A rapid, acetonitrile-free, HPLC method for determination of melamine in infant formula. *Analytica Chimica Acta*, *665*(2), pp.227-230.

Vergani, L., Fanin, M., Martinuzzi, A., Galassi, A., Appi, A., Carrozzo, R., Rosa, M. and Angelini, C., 1990. Liver fatty acid-binding protein in two cases of human lipid storage. *Molecular and cellular biochemistry*, *98*(1-2), pp.225-230.

Verlet, L., 1967. Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical review*, *159*(1), p.98.

Verlet, L., 1968. Computer" experiments" on classical fluids. ii. equilibrium correlation functions. *Physical Review*, *165*(1), p.201.

Vriend, G., 1990. WHAT IF: a molecular modeling and drug design program. Journal of molecular graphics, 8(1), pp.52-56.

Waheed, A.A. and Freed, E.O., 2009. Lipids and membrane microdomains in HIV-1 replication. *Virus research*, *143*(2), pp.162-176.

Wallner, B. and Elofsson, A., 2006. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Science, 15(4), pp.900-913.

Wang, J., Wang, W., Kollman, P.A. and Case, D.A., 2006. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling*, *25*(2), pp.247-260.

Wang, J.K., Kiyokawa, E., Verdin, E. and Trono, D., 2000. The Nef protein of HIV-1 associates with rafts and primes T cells for activation. *Proceedings of the National Academy of Sciences*, *97*(1), pp.394-399.

Wang, Q., Zhuravleva, A. and Gierasch, L.M., 2011. Exploring weak, transient protein–protein interactions in crowded in vivo environments by in-cell nuclear magnetic resonance spectroscopy. *Biochemistry*, *50*(43), pp.9225-9236.

Wang, Y., Li, C. and Pielak, G.J., 2010. Effects of proteins on protein diffusion. *Journal of the American Chemical Society*, *132*(27), pp.9392-9397.

Wang, Y., Liu, J.B., Tang, S.S. and Jin, R.F., 2015. Preparation of melamine molecularly imprinted polymer by computer-aided design. *Journal of separation science*, *38*(15), pp.2647-2654.

Wang, Y., Sarkar, M., Smith, A.E., Krois, A.S. and Pielak, G.J., 2012. Macromolecular crowding and protein stability. *Journal of the American Chemical Society*, *134*(40), pp.16614-16618.

Watanabe, H., Shiratori, T., Shoji, H., Miyatake, S., Okazaki, Y., Ikuta, K., Sato, T. and Saito, T., 1997. A novel acyl-CoA thioesterase enhances its enzymatic activity by direct binding with HIV Nef. *Biochemical and biophysical research communications*, *238*(1), pp.234-239.

Watson, J.D. and Crick, F.H., 1953. Molecular structure of nucleic acids. *Nature*, *171*(4356), pp.737-738.

Webb, B. and Sali, A., 2014. Comparative protein structure modeling using Modeller. Current protocols in bioinformatics, pp.5-6.

Whitcombe, M.J., Rodriguez, M.E., Villar, P. and Vulfson, E.N., 1995. A new method for the introduction of recognition site functionality into polymers prepared by molecular imprinting: synthesis and characterization of polymeric receptors for cholesterol. *Journal of the American Chemical Society*, *117*(27), pp.7105-7111.

Wierzbicki, A.S., 2012. New directions in cardiovascular risk assessment: the role of secondary risk stratification markers. *International journal of clinical practice*, *66*(7), pp.622-630.

Williams, D.H., Cox, J.P., Doig, A.J., Gardner, M., Gerhard, U., Kaye, P.T., Lal, A.R., Nicholls, I.A., Salter, C.J. and Mitchell, R.C., 1991. Toward the semiquantitative estimation of binding constants. Guides for peptide-peptide binding in aqueous solution. *Journal of the American Chemical Society*, *113*(18), pp.7020-7030.

Williams, E.L., Acquaviva, C., Amoroso, A., Chevalier, F., Coulter-Mackie, M., Monico, C.G., Giachino, D., Owen, T., Robbiano, A., Salido, E. and Waterham, H., 2009. MUTATION UPDATE. *Hum Mutat*, *30*, pp.910-917.

Wuff, G. and Sarhan, A., 1972. The use of polymers with enzyme-analogous structures for the resolution of racemate. *Journal of the Angewandte Chemie International Edition*, *11*(3), pp.341-345.

Xing, Q., Huang, P., Yang, J., Sun, J.Q., Gong, Z., Dong, X., Guo, D.C., Chen, S.M., Yang, Y.H., Wang, Y. and Yang, M.H., 2014. Visualizing an Ultra-Weak Protein–Protein Interaction in Phosphorylation Signaling. *Angewandte Chemie*, *126*(43), pp.11685-11689.

Yan, H. and Row, K.H., 2006. Characteristic and synthetic approach of molecularly imprinted polymer. *International journal of molecular Sciences*, *7*(5), pp.155-178.

Yang, A.S. and Honig, B., 1999. Sequence to structure alignment in comparative modeling using PrISM. Proteins: Structure, Function, and Bioinformatics, 37(S3), pp.66-72.

Yarnitzky, T., Levit, A. and Niv, M.Y., 2010. Homology modeling of G-protein-coupled receptors with X-ray structures on the rise. Curr Opin Drug Discov Devel, 13(3), pp.317-25.

Yonath, A. 2010. Polar Bears, Antibiotics, and the Evolving Ribosome (Nobel Lecture). Wiley-VCH Verlag.

York, D.M., Wlodawer, A., Pedersen, L.G. and Darden, T.A., 1994. Atomic-level accuracy in simulations of large protein crystals. *Proceedings of the National Academy of Sciences*, *91*(18), pp.8715-8718.

York, D.M., Wlodawer, A., Pedersen, L.G. and Darden, T.A., 1994. Atomic-level accuracy in simulations of large protein crystals. *Proceedings of the National Academy of Sciences*, *91*(18), pp.8715-8718.

Zhang, X., Roe, S.M., Hou, Y., Bartlam, M., Rao, Z., Pearl, L.H. and Danpure, C.J., 2003. Crystal structure of alanine: glyoxylate aminotransferase and the relationship between genotype and enzymatic phenotype in primary hyperoxaluria type 1. *Journal of molecular biology*, *331*(3), pp.643-652.

Zhou, H. and Zhou, Y., 2005. SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. Bioinformatics, 21(18), pp.3615-3621.

Zhou, H.X., Rivas, G. and Minton, A.P., 2008. Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annual review of biophysics*, *37*, p.375.

Zipeto, D. and Beretta, A., 2012. HLA-C and HIV-1: friends or foes?. *Retrovirology*, *9*(1), p.1.

Web references:

AutoDockTools. http://autodock.scripps.edu/resources/adt (accesses 12.10.2015)

Autodock parameters. http://autodock.scripps.edu/resources/parameters (accesses 20.10.2015)

MODELER program. http://salilab.org/modeller/ (accesses 20.07.2015)

MARTINI force field. http://md.chem.rug.nl/ (accesses 08.09.2015)

Autodock Vina manual. http://vina.scripps.edu/manual.html (accesses 12.10.2015)

Bulk Polymerization. http://digilander.libero.it/Imprinted_Polymers/MIPs/sintesi_bulk.htm (accesses 23.9.2014)

Precipitation Polymerization.
http://digilander.libero.it/Imprinted_Polymers/MIPs/sintesi_precipitazione.htm (accesses 23.9.2014)

Grafting polymerization on monodisperse polymer beads.
http://digilander.libero.it/Imprinted_Polymers/MIPs/sintesi_grafting.htm (accesses 23.9.2014)

Emulsion synthesis. http://digilander.libero.it/Imprinted_Polymers/MIPs/sintesi_emulsione.htm (accesses 23.9.2014)

Link to MIRATE web server. http://mirate.di.univr.it/ (accesses 20.11.2016)

Documentation and Tutorial of MIRATE web server step 1 and 2.
http://mirate.di.univr.it/dock/docs.php (accesses 20.11.2016)

Documentation and Tutorial MIRATE web server step 3. http://mirate.di.univr.it/dyn/docs.php (accesses 20.11.2016)

Appendix:

Chapter 4

Legends:

S: Single Haddock docking; M: Multiple Haddock docking; V: VINA pipeline; A: Autodock pipeline; Y: Yes; N: No.

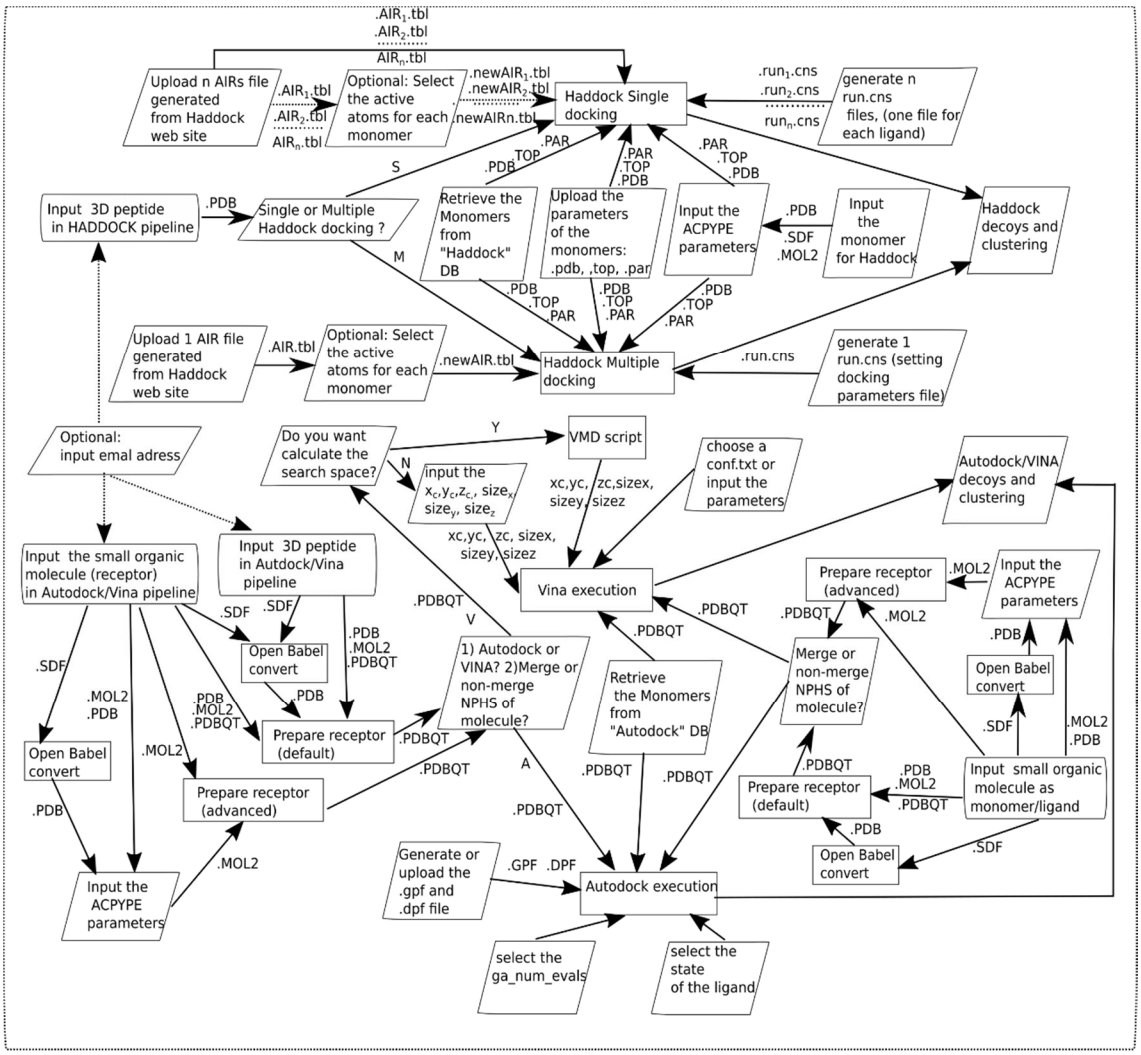


Figure S4.1.a. A detailed workflow for step1 (input preparation) and step2 (docking step) in MIRATE web server.

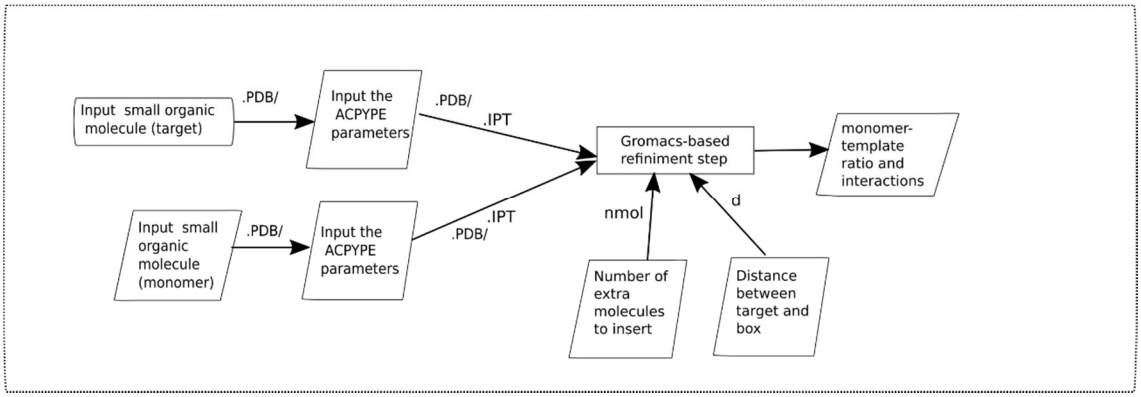


Figure S4.1.b. A detailed workflow for step 3 (refinement step) in MIRATE web server.

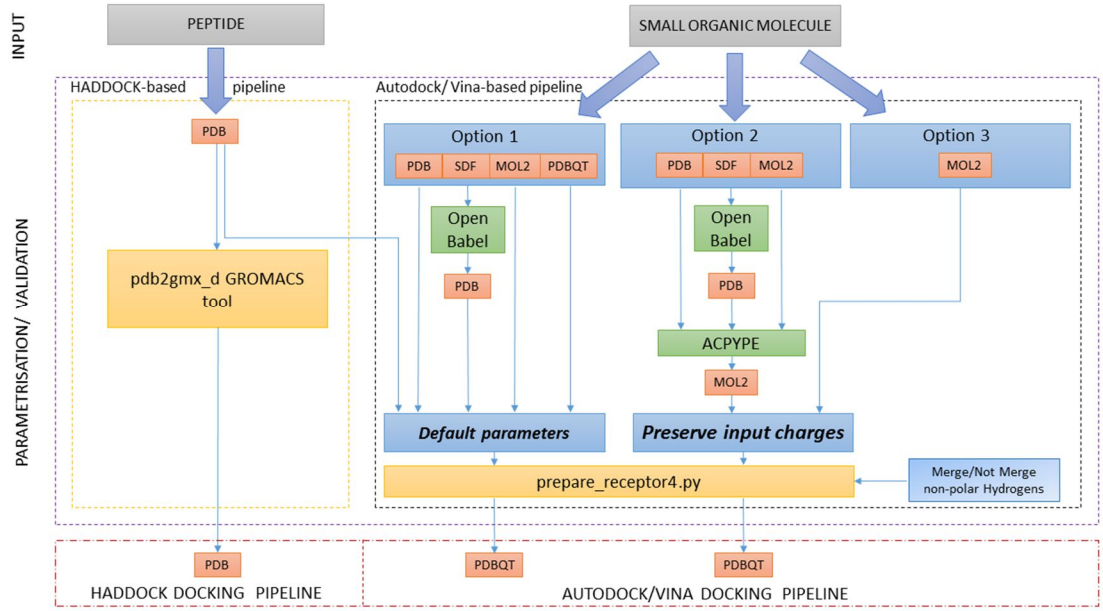


Figure S4.2. A detailed workflow of the Receptor preparation step in MIRATE web server.

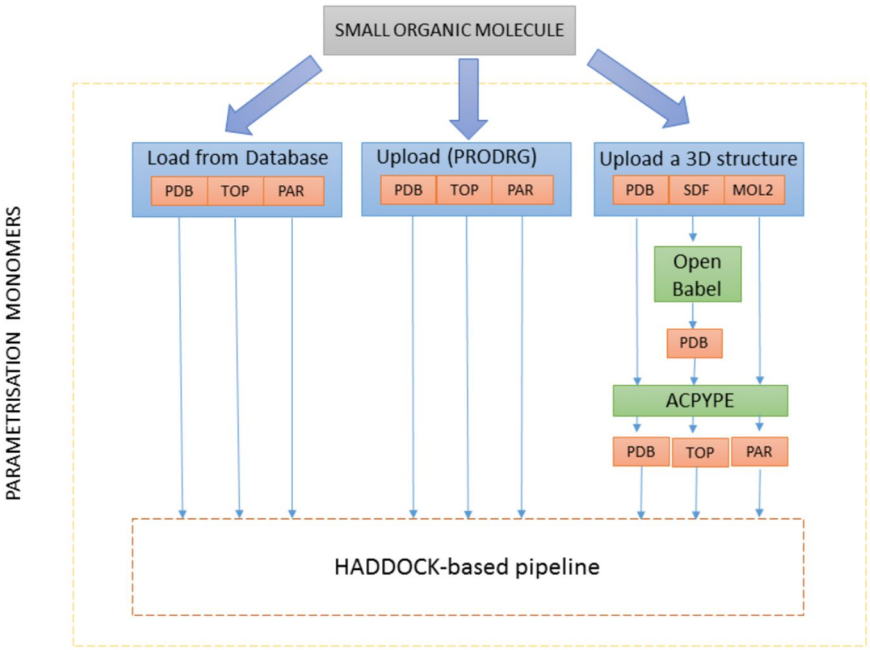


Figure S4.3. A detailed workflow of the Monomer preparation for HADDOCK pipeline in MIRATE web server.

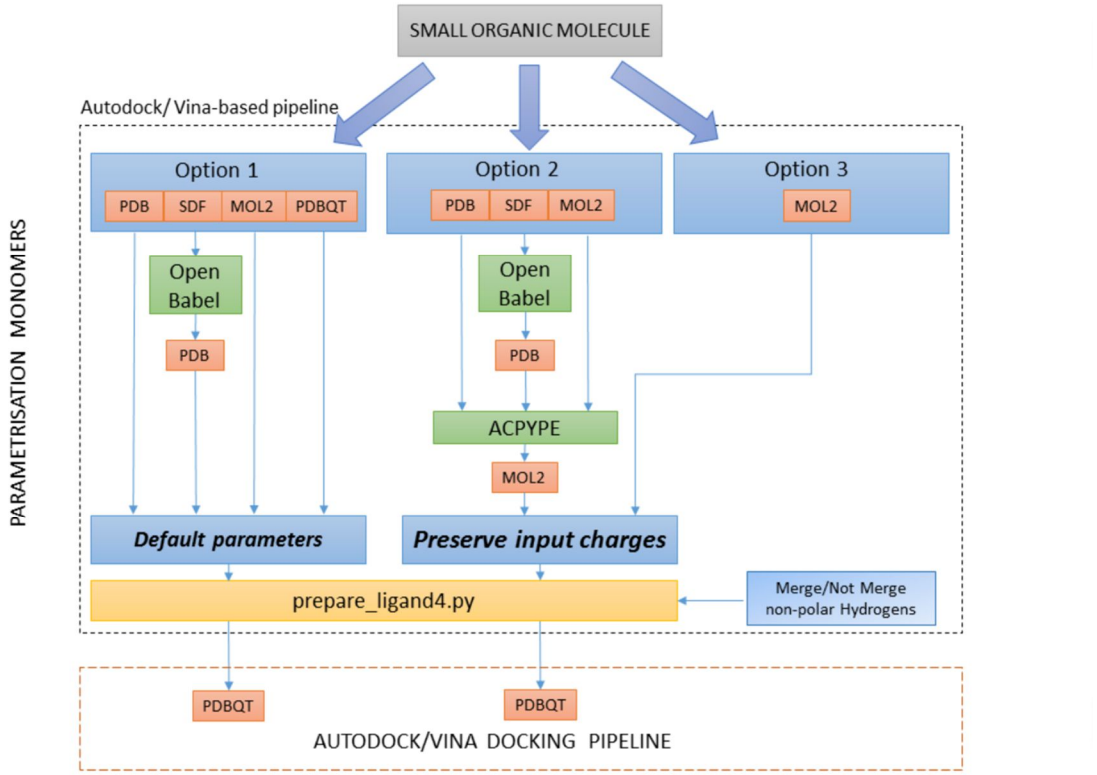


Figure S4.4. A detailed workflow of the Monomer preparation for Vina/Autodock pipeline in MIRATE web server.

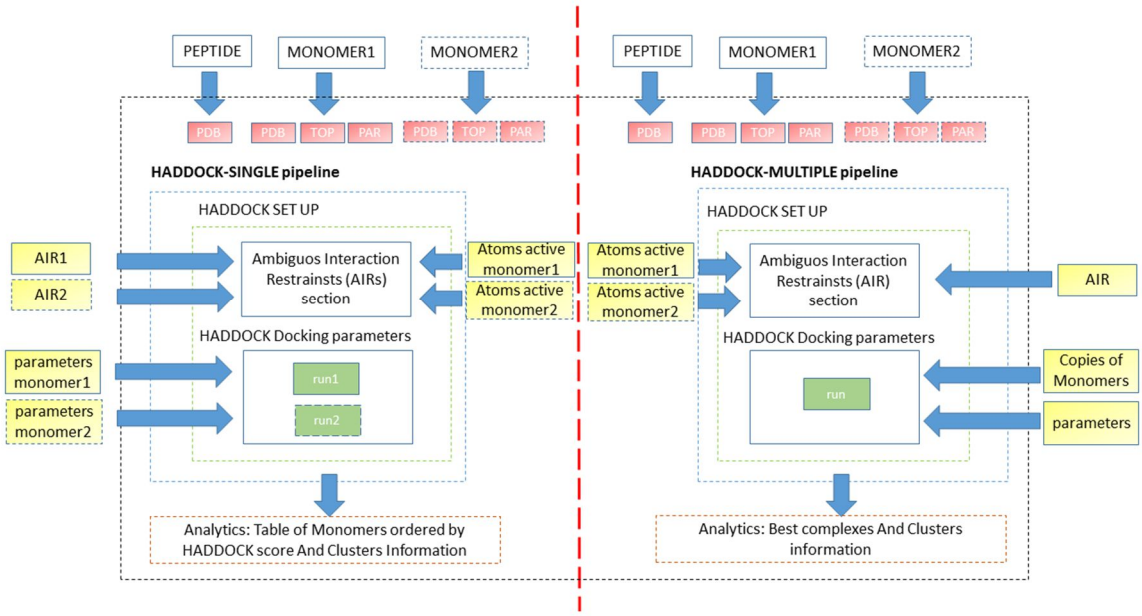


Figure S4.5. A detailed workflow of the HADDOCK docking pipeline in MIRATE web server.

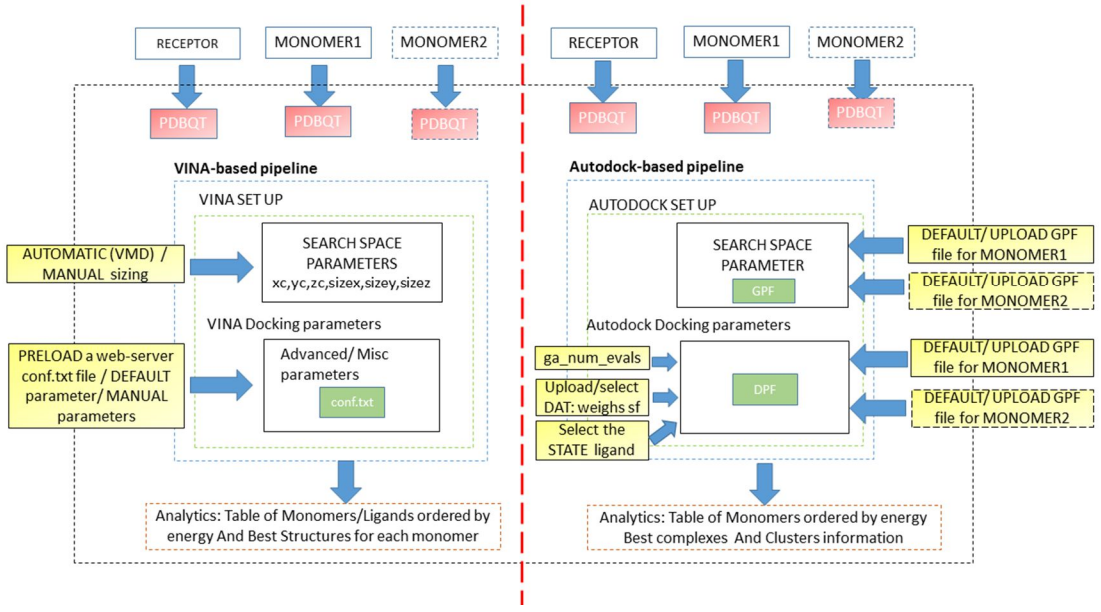


Figure S4.6. A detailed workflow of the Vina and Autodock docking pipelines in MIRATE web server.

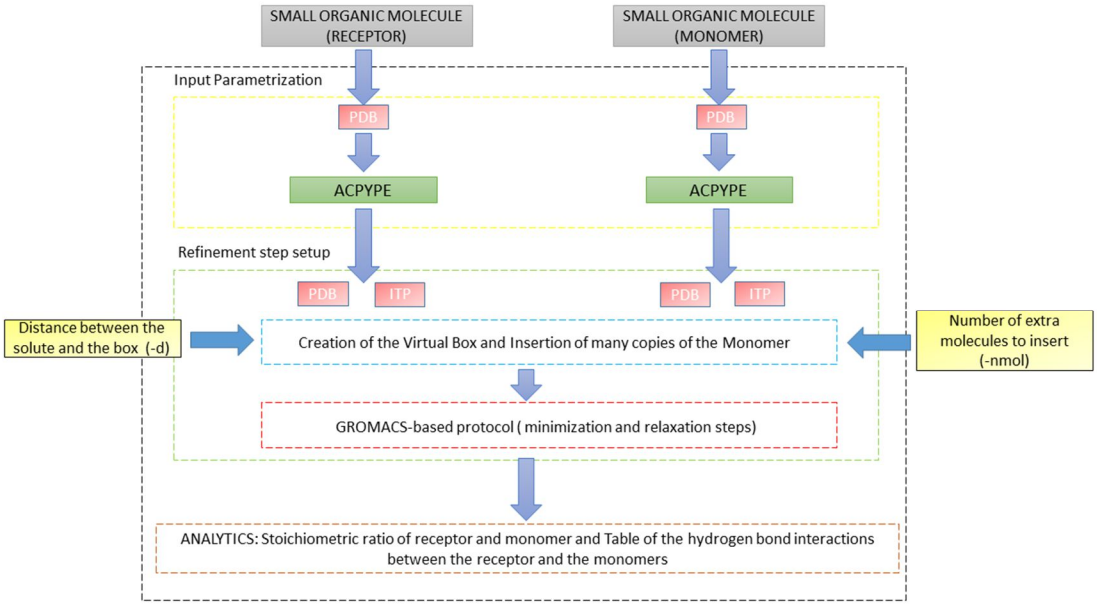


Figure S4.7. A detailed workflow for the Refinement step in MIRATE web server.

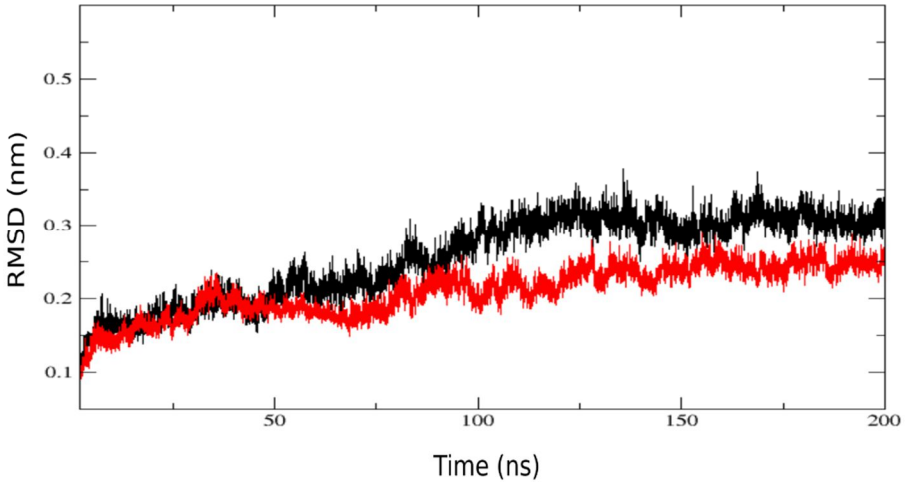


Figure S6.1. Plot of the RMSD vs time. The simulations were performed on the AGT-Mi (black) and I244T (red) dimeric systems. Reprinted from Dindo et al., 2016.

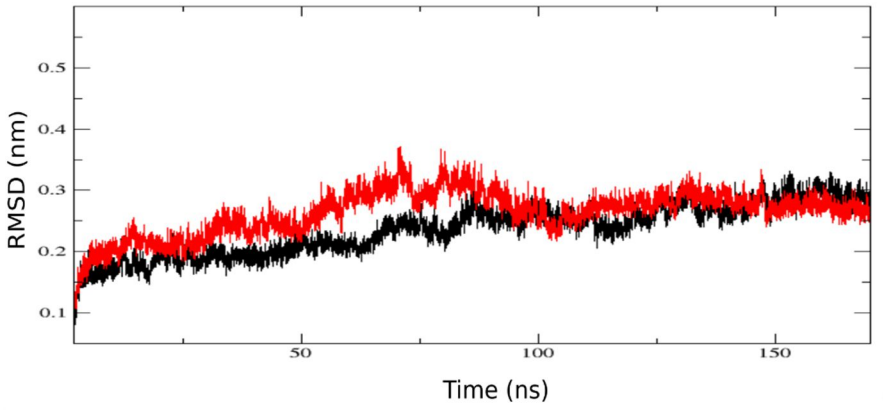


Figure S6.2. Plot of the RMSD vs time. The simulations were performed on the AGT-Mi (black) and F152I (red) monomeric systems. Reprinted from Dindo et al., 2016.

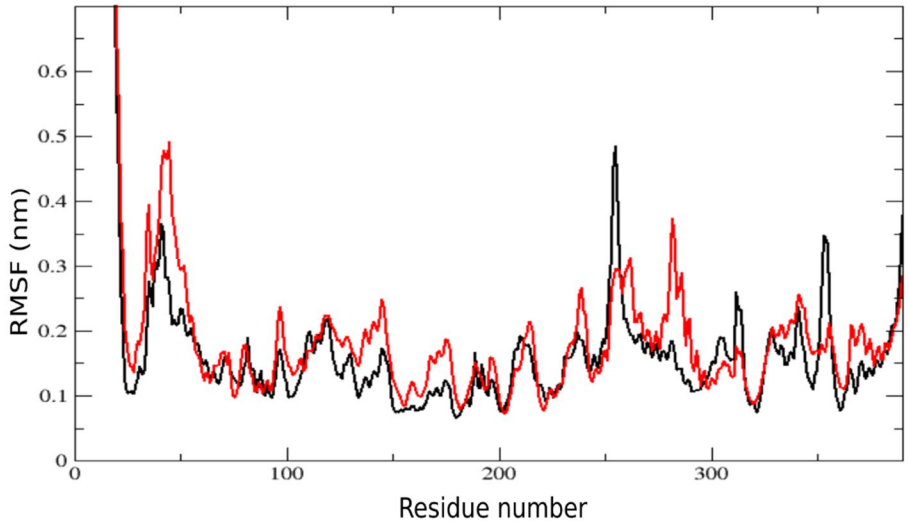


Figure S6.3. RMSF plot. The simulations were performed on the AGT-Mi (black) and F152I (red line) monomeric systems. Reprinted from Dindo et al., 2016.

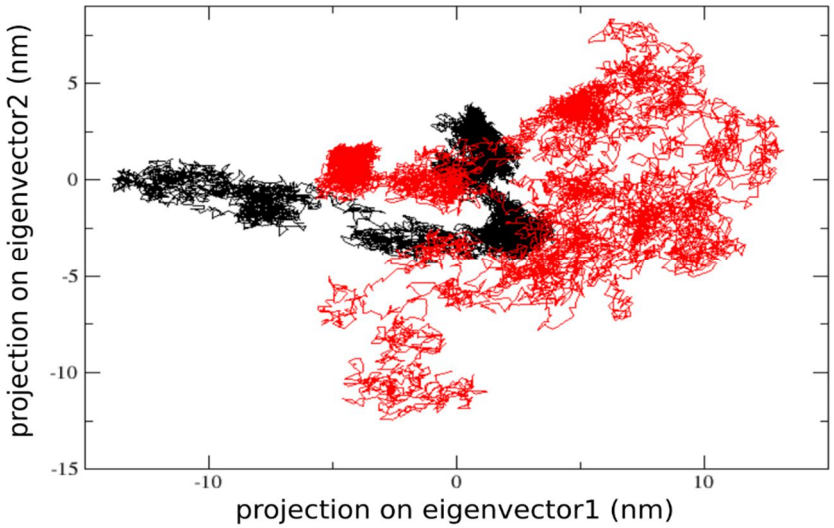


Figure S6.4. 2D projection of AGT-Mi (black) and F152I-Mi mutant (red) monomer models over the first two principal components. The AGT-Mi (black) form covers a smaller region of conformational space than the F152I-Mi (red) mutant form. Reprinted from Dindo et al., 2016.

Chapter7

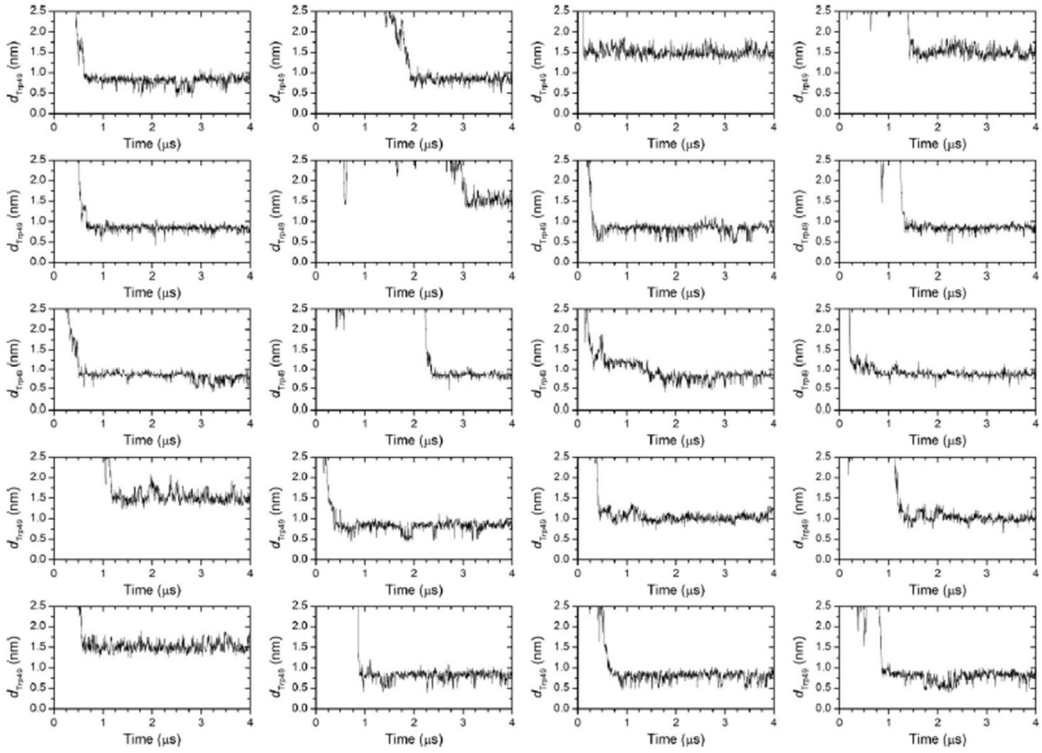


Figure S7.1. Time evolution of dTrp49 along twenty CD MD simulations at low salt concentration (0.030 M NaCl). For the sake of clarity, distance (dTrp49) values below 2.5 nm are shown. Reprinted from Ceccon et.al., 2015.

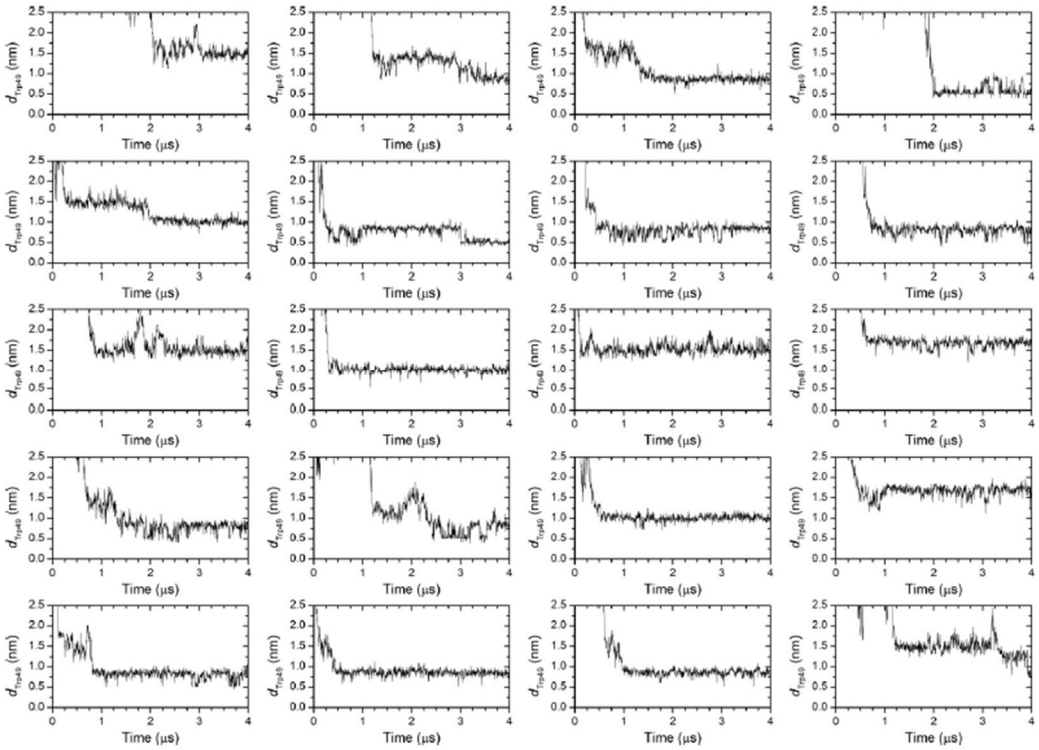


Figure S7.2. Time evolution of dTrp49 along twenty CD MD simulations at high salt concentration (0.175 M NaCl). For the sake of clarity, distance (dTrp49) values below 2.5 nm are shown. Reprinted from Ceccon et.al., 2015.

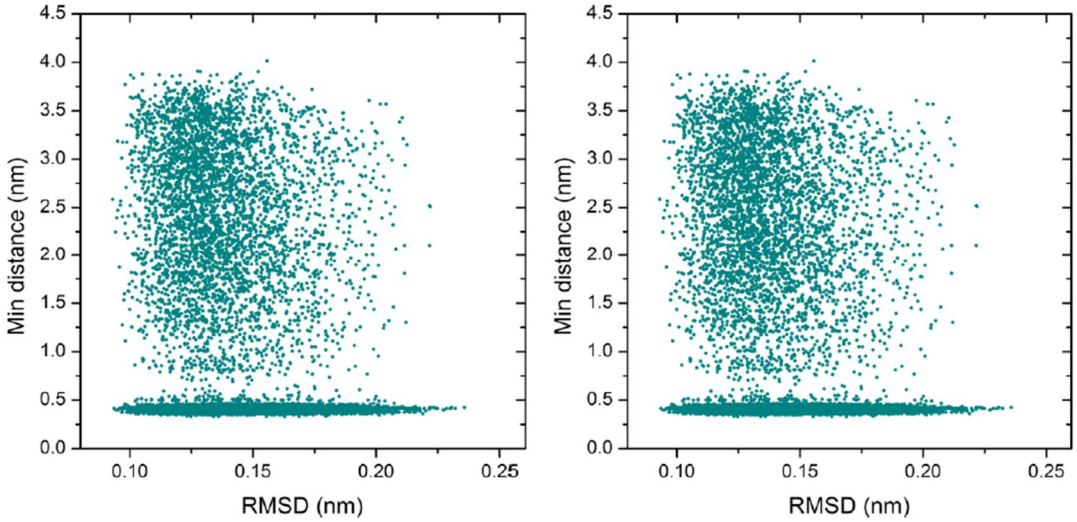


Figure S7.3. Minimum distance between IBABP and the POPG bilayer vs. root mean square deviation (RMSD) of the backbone CG beads of IBABP for the CG MD simulations performed at low (left) and high (right) ionic strength. Reprinted from Ceccon et.al., 2015.

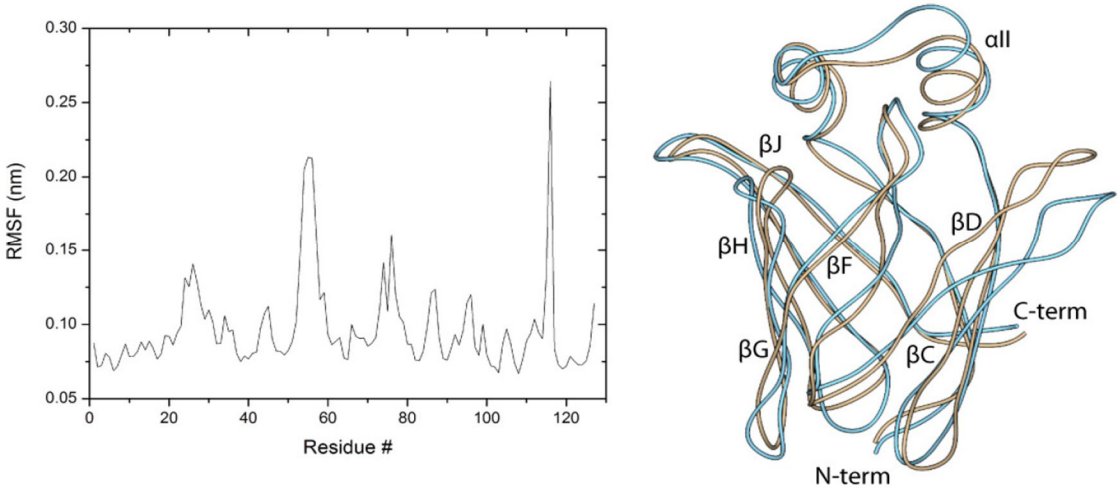


Figure S7.4. (Left) Root mean square fluctuations of IBABP calculated between selected frames of the CG MD simulations of the protein in solution and in contact with the membrane. (Right) Superimposition of the IBABP backbone in solution (light blue) and in contact with the membrane (light brown). The α-helix and the β-strands discussed in the main text are indicated. Reprinted from Ceccon et.al., 2015.